Review of Education
An International Journal of Major Studies in Education
BERA BRITISH EDUCATIONAL RESEARCH ASSOCIATION
Check for updates

Review of Education
Vol. 9, No. 3, October 2021, e3270

DOI: 10.1002/rev3.3270

PISA 2018 in England, Northern Ireland, Scotland and Wales: Is the data really representative of all four corners of the UK?

John Jerrim*
UCL Social Research Institute, UK

PISA is an influential international study of the achievement of 15-year-olds. It has a high profile across the devolved nations of the UK, with the results having a substantial impact upon education policy. Yet many of the technical details underpinning PISA remain poorly understood—particularly amongst non-specialists—including important nuances surrounding the representivity of the data. This paper provides new evidence on this issue, based upon a case study of PISA 2018. I illustrate how there are many anomalies with the data, with the combination of nonresponse, exclusions from the test and technical details surrounding eligibility criteria leading to total nonparticipation rates of around 40% (amongst the highest anywhere in the world). It is then shown how this leads to substantial uncertainty surrounding the PISA results, with clear evidence of bias in the sample for certain parts of the UK. I conclude by discussing how more transparent reporting of the technical details underpinning PISA is needed, at both a national and international level.

Keywords international comparisons, PISA, sample selection.

Introduction

The Programme for International Student Assessment (PISA) is an influential international study of 15-year-olds' ability to use their reading, mathematics and science knowledge and skills to meet real-life challenges. Conducted every three years since 2000, it has become a widely-watched indicator of national educational performance across the globe. Results from PISA have had substantial real-world impact upon education policy (Baird *et al.*, 2011). This includes reforms made to national curricula in South Korea and Mexico, along with alterations to national assessments in Slovakia and Japan (Breakspear, 2012). Policy recommendations made by the Organisation for Economic Cooperation and Development (OECD) off the back of the PISA results have also been influential in Wales (OECD, 2014) and a wide-range of middle income countries (Lockheed *et al.*, 2015), along with many other international examples. It is now one of the most influential studies in education, with the triannual results impacting upon the thoughts and actions of key decisions made all around the world.

PISA has also had a notable impact upon discussion and debates on education in the United Kingdom—the country of focus in this paper. Since devolution in the late

*Correspondence: UCL Social Research Institute, University College London, 20 Bedford Way, London WC1H 0AL, UK. Email: J.Jerrim@ucl.ac.uk

1990s, education policies, practices and qualifications have diverged across England, Northern Ireland, Scotland and Wales. This has led to questions about how the four nations of the UK compare in terms of young people's educational achievement, and how this has changed over time (Machin *et al.*, 2013). With few other comparable sources of data available, PISA has become the 'go-to' resource to conduct comparisons of educational achievement across the UK. Indeed, national reporting of each new round of PISA has an entire chapter devoted to intra-UK comparisons of educational performance (Sizmur *et al.*, 2019a,b,c), with these results then widely reported within the national media (Coughlan, 2019). For reference, the latest trends in PISA mathematics scores can be found in Figure 1.

Given PISA's now prominent role in our understanding of educational performance across the UK, it is vital that it provides sound and reliable evidence upon which such comparisons can be made. Yet some have questioned various aspects surrounding the reliability of the PISA study, both within the UK and internationally. For instance, investigating trends in England's PISA scores over time, I noted in Jerrim (2013) how many important changes were made between PISA 2000/2003 and subsequent rounds, affecting both response rates and test month, which may then have impacted upon the results. In the case of Turkey, Spaull (2019) noted how nuances surrounding the PISA eligibility criteria are likely to have a big impact upon the reliability of trends in PISA scores over time. Anders *et al.* (2021) conducted a detailed case study of the PISA 2015 data for Canada, illustrating how a combination of low response rates and high exclusions lead to serious questions surrounding the representivity of the data. In the case of Portugal, Pereira (2011) argued that changes to how the sample was drawn had a substantial impact upon changes in the PISA scores over time. Also in Portugal, Freitas *et al.* (2016) found non-trivial differences between the PISA target population and the final sample, with this then having a

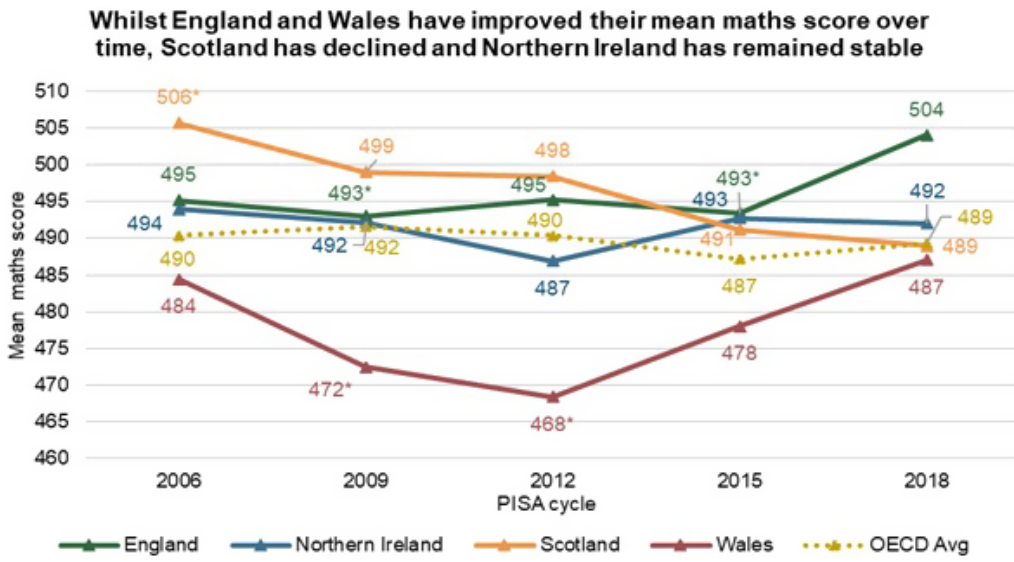

Figure 1. Trends in PISA mathematics scores across the UK. Source: Sizmur *et al.* (2019a,b,c).

notable impact upon changes in PISA scores. Concerns have also been raised regarding the switch between paper and computer assessment in PISA, which occurred in 2015 (Jerrim, 2016; Jerrim *et al.*, 2018), and how this may have affected comparisons of results across countries and over time. Other issues have been raised about the lack of transparency over the item-response theory model used to generate the PISA scores, including a lack of transparency about how the so-called 'plausible values' are produced (Goldstein, 2017). For instance, Zieger *et al.* (2020) illustrated how subtle changes made to the PISA scaling model can have a big impact upon cross-national comparisons of educational inequality. More generally, questions have been raised surrounding cross-country differences in translation and interpretation of the PISA test material (Kankaraš & Moors, 2014; El Masri *et al.*, 2016).

Much of the aforementioned work serves as a platform for this paper. After the UK was disqualified from PISA in 2003—due to its low response rate—significant efforts were made to ensure that data collected in future waves would be more robust. This included moving the test date in England, Wales and Northern Ireland to avoid clashes with GCSE preparation, the appointment of external contractors to collect the data rather than a government department, and introducing legislation that could potentially force schools to participate if they did not willingly comply. Taken at face value, it may seem that this strategy has been successful; no part of the UK was excluded from PISA in the 2006, 2009, 2012, 2015 or 2018 rounds of the study, with the data always being considered by the OECD to be of acceptable quality.

However, in reality, the true situation remains much more nuanced than first meets the eye. As this paper will explain, there continues to be a great amount of nonparticipation by pupils and schools in PISA across the UK, leading to potential biases in the data. I pay particular attention to the PISA 2018 data for Scotland, where particular complications and anomalies have emerged. In summary, my discussion will illustrate how:

- The nonparticipation rate in PISA across the UK (and in Scotland as an individual entity) is around 40%. This is amongst the highest anywhere in the world.
- This high-level of nonparticipation means that there is a large amount of uncertainty surrounding the UK's PISA results. This is likely to affect the reliability of comparisons that can be made across the four UK nations, comparisons with other countries and how results have changed over time.
- Key issues regarding data quality and comparability have—in my view—not been adequately reported, with greater transparency needed in future rounds.
- There is clear evidence of an upward bias in the PISA 2018 data for England and Wales, with lower-achievers systematically excluded from the sample.
- If a truly representative sample of the population had taken the tests, average PISA scores in England and Wales would likely be around 10–15 points lower.

The main aim of this investigation is to help a broader group of interested stakeholders understand such key issues, and to aid their interpretation of the PISA data for the UK. Yet it also highlights a need for better reporting practices of the PISA results in the future, both within the UK and by the OECD. I thus conclude by calling upon the UK Statistics Authority to conduct a review of the PISA 2018 data for the

UK, and for them to issue some 'best practice' guidelines for the reporting of data from future PISA waves.

The paper now proceeds as follows. Background to how the PISA sample for the UK is drawn provides context to the design of PISA and how it is implemented across the UK. Anomalies in the PISA 2018 data for Scotland focuses upon issues with the PISA 2018 data for Scotland, while Anomalies in the PISA 2018 data for the rest of the UK provides an analogous discussion for England, Northern Ireland and Wales. Conclusions and recommendations for future PISA data collections follow in Conclusions.

Background to how the PISA sample for the UK is drawn

Background

The OECD—who lead the PISA study—treat the UK as a single country (Sizmur *et al.*, 2019a). This means that it is the data for the UK as a whole that is subject to the OECD's 'technical standards'. Out of the four UK nations, only Scotland participates in PISA as an 'adjudicated' sub-national entity (i.e. a fully-fledged, stand-alone participant). This means that additional technical details are reported for Scotland in the annexes to the OECD's PISA technical reports (OECD, 2019). As a sub-national entity, Scotland is also held accountable (as an individual entity) to the OECD's technical standards.

Although England, Wales and Northern Ireland do not participate in PISA as adjudicated sub-national entities (and are not individually judged against the OECD's technical standards) they do draw an oversample of schools to facilitate national reporting. Thus each of the four UK nations produce their own national analyses (Scottish Government, 2019; Sizmur *et al.*, 2019a,b,c), with separate figures for England, Wales, Northern Ireland and Scotland reported on the PISA results day. Exactly how the UK (and its four constituent nations) participates in PISA is therefore somewhat more complicated than first meets the eye.

Target population

PISA is widely interpreted as a measure of 15-year-olds skills' in science, reading and mathematics. However, the actual target population is somewhat more nuanced, defined as 'students aged between 15 years and three (completed) months and 16 years and two (completed) months at the beginning of the period of testing, attending educational institutions located within the adjudicated entity, and in Grade 7 or higher' (OECD, 2019: Annex I). As noted previously (Spaull, 2019), the specifics underpinning this definition—particularly the focus upon pupils who are enrolled in school—has some important implications. In particular, it is likely to inflate PISA scores in countries where a non-trivial proportion of 15-year-olds are not enrolled in school (mainly lower and middle income settings).

Sampling frame and school-level exclusions

With the target population in hand, a sampling frame is constructed—essentially a list of all schools within a country which includes 15-year-old pupils. However, from this sampling frame, countries are permitted to exclude some schools due to either logistical reasons or where there is an expectation that most pupils would not be eligible to participate (OECD, 2019). In England, for example, special schools, hospital schools, secure units, international immersion schools and pupil referral units were excluded on this basis (Sizmur *et al.*, 2019a). The OECD data quality standards stipulate that a maximum of 2.5% of schools can be excluded for such reasons (OECD, 2019: Annex I)[1], with the PISA 2018 data for the UK within this limit (2.2% for the UK and 1.7% for Scotland). Nevertheless, any such school-level exclusions made by a country could contribute to the PISA data becoming unrepresentative of the target population.

School sampling

After excluding a small number of schools, those remaining on the sampling frame within each country are 'stratified' into different groups (known as explicit stratification). The precise stratification variables used within each of the four UK nations differ (see Appendix 1 for details) but typically include some combination of broad geographic region and school type. Then, within these explicit strata, schools are ranked/ordered by a set of further characteristics (known as implicit stratification). The most important stratification variable used in England, Scotland and Wales is historic performance in national examinations (e.g., Attainment 8 scores in England)[2]. Within each of the four UK countries—and within each explicit strata—schools are then sampled with probability proportional to size.

School nonresponse

As with any study, not all the schools that are asked to participate in PISA agree to do so; there is a problem of school nonresponse. A somewhat unusual feature of PISA is that, if a school refuses to participate, a substitute/replacement can take its place. Specifically, for each school initially sampled, two possible replacements are also selected at the same time. These are typically schools that are adjacent to the originally sampled schools on the sampling frame, and should thus be similar in terms of historic school performance on national examinations (at least in England, Wales and Scotland, where this information is used in the stratification of the sample). In reality, this approach to school nonresponse is a form of imputation, with an implicit Missing At Random (MAR) assumption being made.

The OECD set criteria for the level of school nonresponse they deem 'acceptable', as illustrated by Figure 2. The aim is for each country to successfully recruit 85% of originally sampled schools (before any replacements are included), with the vast majority achieving this in PISA 2018. In contrast, if less than 65% of originally sampled schools fail to participate, then the data for the country should be considered of unacceptable quality and excluded from the PISA results (although, in reality, there

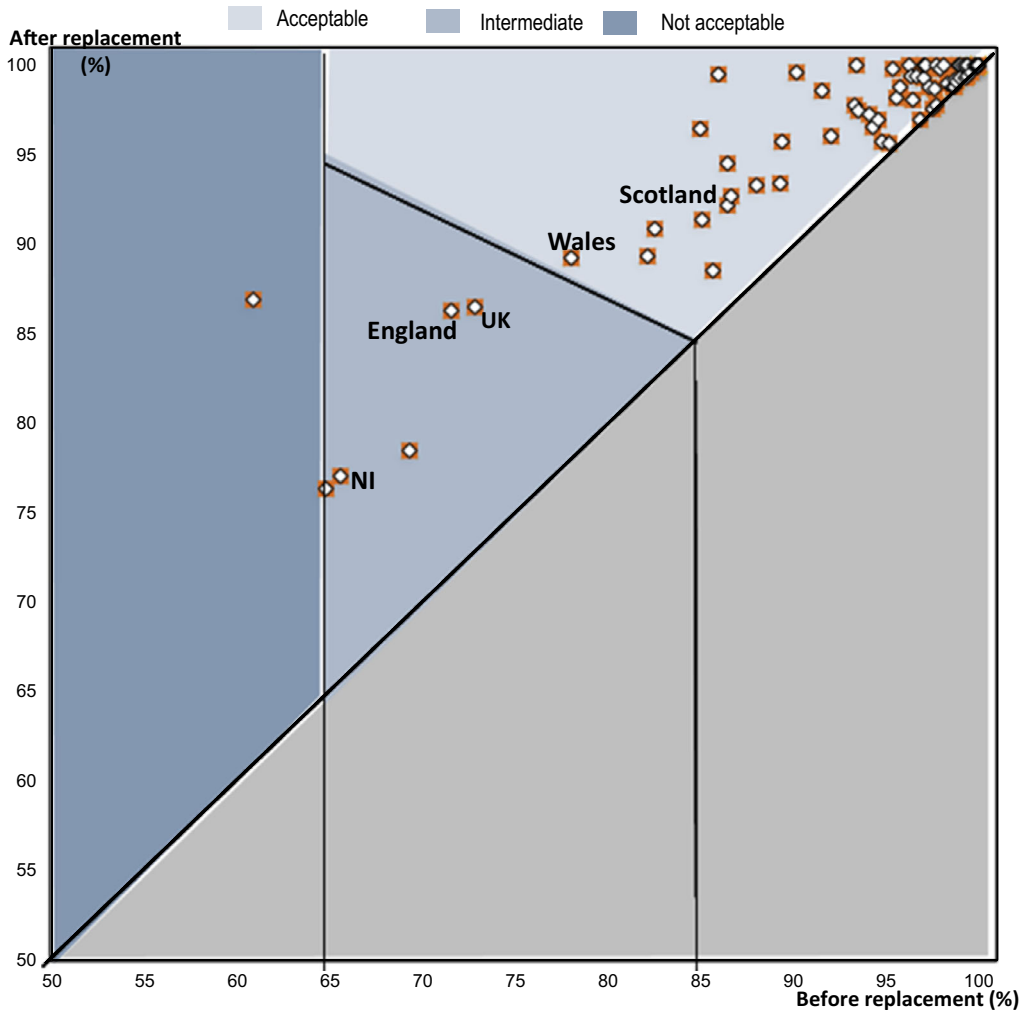

Figure 2. School response rates before and after replacement in PISA 2018.Notes: Horizontal axis refer to the 'before replacement' school response rate—the percentage of initially sampled schools that completed the PISA test. Figures on the vertical axis refer to the 'after replacement' response rate—the percentage of schools that completed the PISA test after substitute schools are included in the figures. The dark-blue area, where the 'before replacement' level is below 65%, means the technical standard has not been met and the country should be excluded for the PISA study. The light blue 'acceptable' area is where countries are fully compliant with the PISA school response rate technical standard. The 'intermediate zone' in the middle refers to where the OECD technical standard has not been fully met, with countries required to complete a school-level NRBA.

are some countries that fail to reach this 'minimum' benchmark and are not excluded by the OECD—see Anders *et al.*, 2021). If a country achieves between a 65% and 85% response rate amongst initially sampled schools, then replacement schools can be included to meet the OECD's school response rate criteria. However, the school response rate target after replacement also increases. For instance, if a country has a 70% response rate amongst initially sampled schools, they would need to achieve a

93% school response rate after the replacements are included in order to fulfil the OECD's technical standards (Sizmur *et al.*, 2019a,b,c). If they fail to do so, then a country must produce a school-level nonresponse bias analysis (NRBA) to demonstrate whether there is any bias in the final school sample. This NRBA is adjudicated by the OECD to decide whether to include the country's data in the PISA results. However, as noted by Anders *et al.* (2021), this adjudication process is actually quite weak, with only three out of 23 instances of a NRBA leading to a country being excluded between PISA 2000 and 2015.

Importantly, the PISA 2018 data for the UK 'did not fully meet the PISA 2018 participation requirement' (Sizmur *et al.*, 2019a,b,c) due to high levels of school nonresponse—as illustrated by Figure 2. School nonresponse was a particular problem in England (school response rate 72% before replacement) and Northern Ireland (66%), while Wales and Scotland met the OECD's criteria. The OECD thus required the UK (as a whole, single entity) to produce a NRBA. The OECD's technical group judged 'that no notable bias would result from the [school] nonresponse' (OECD, 2019: Chapter 14). I return to this point in Anomalies in the PISA 2018 data for the rest of the UK, when discussing issues found with the PISA 2018 data for England and Northern Ireland.

Within-school sampling of pupils

All schools that agree to participate in PISA are asked to provide a list of all pupils who meet the definition of the PISA target population (i.e. pupils aged 15 years and three months and 16 years and two months) at the time the assessment is due to take place. Using these lists, 40 pupils are randomly selected from within each school to participate in PISA. Note that the age-based definition used in PISA means that the pupils selected may fall across multiple school year groups (an important point which will be returned to when discussing anomalies in the Scottish PISA data in the section that follows).

However, not all of these age-eligible pupils who have been selected to take the PISA test will actually sit the assessment. In this paper, we term this 'nonparticipation', noting that this can occur for three reasons.

The first is 'within-school' exclusions—meaning that schools can decide not to test some of the sampled pupils. The OECD technical standards state that such within-school exclusions should total less than 2.5% of the PISA-desired target population, and that the combination of school-level and within-school exclusions should not exceed 5% of the target population (OECD, 2019: Annex I).

The second is ineligibility: pupils who were included on the age-eligible pupil list, but who were then considered to not meet the definition of the target population. Importantly, this ineligibility category includes pupils who left the school in between the time the sample was drawn and the time the assessment was conducted (McKeown *et al.*, 2019). Specifically, the PISA 2018 report for Scotland notes how *students that had left the school in the interim* [between the time of the pupil sample being drawn and the time of the test] *were not considered part of the target sample* (Scottish Government, 2019). As far as I am aware, the OECD's technical standard and data quality criteria set no maximum limits on the percentage of pupils deemed ineligible.

Finally, there is the issue of pupil nonresponse. The remaining pupils (or their parents) may not consent to take part in the study or pupils may be absent on the day of the test. The OECD technical standards stipulate that such cases of pupil nonresponse must not be greater than 20%, otherwise a pupil-level NRBA will need to be conducted. In reality, almost all countries meet these standards (e.g., in PISA 2018, out of the 80 participating countries, just one—Portugal—did not meet this threshold), potentially illustrating how this may not be considered a particularly high stretch target to meet.

A note about weights

The OECD database includes a set of weights. Given the issues discussed above, it is important to understand what these weights achieve, and the implications for analysis of PISA data for the UK.

The first key function of these weights is that they correct estimates for unequal probabilities of schools being selected into the PISA sample (in part due to the oversampling that occurs across the UK). This element of the weights also scales figures up to the UK population. A key implication is that all figures reported for the UK by the OECD are driven by the data for England, given that this country accounts for 84% of the 15-year-olds who live in the UK (Sizmur *et al.*, 2019c). This, in-turn, also means that the UK-wide figures reported by the OECD serve as a close proxy of the results for England as a stand-alone country. On the other hand, the UK-wide figures almost completely mask the situation in Scotland, Northern Ireland and Wales.

The second key role of these weights is that they make some limited adjustment for nonresponse. Specifically, the weights use school-level data in the form of the stratification variables (see Appendix 1), along with some very basic pupil characteristics (year group and grade) to try and account for school and pupil nonresponse. As the stratification variables for England, Wales and Scotland include measures of historical school performance in national examinations, these weights may adequately correct for school nonresponse (see Micklewright *et al.*, 2012 for some empirical evidence on this issue). On the other hand, as argued by Anders *et al.* (2021), the weights provided are highly unlikely to reduce bias due to pupil nonparticipation, given the very limited amount of pupil-level data (just gender and grade) included in their construction. Moreover, previous work has suggested that it is nonresponse amongst pupils—rather than by schools—that drove bias in the PISA data for England in PISA 2000 and 2003 (Micklewright *et al.*, 2012). Thus, in reality, the weighting scheme used within PISA is unlikely to solve potential bias induced by the various ways pupils drop out of the study (particularly when these are not due to school nonresponse).

Test month

A final unusual feature of PISA in the UK is when the assessment takes place. In most northern hemisphere countries, PISA is conducted between March and August. However—since 2006—England, Wales and Northern Ireland have received special dispensation from the OECD to conduct PISA between October and December, so as to avoid conflicts with GCSE examinations[3]. One important implication of this is

that almost all pupils who sit the PISA test in England (97%), Wales (98%) and Northern Ireland (92%) are in the equivalent of Year 11.

In Scotland, the situation has been different. Up until 2015, the PISA test was conducted between March and May. This changed, however, in 2018 when the test period moved to between October and December. The reason behind making this change has not—to my knowledge—been documented either in the Scottish or OECD-reporting of the PISA results. It may however, as I will discuss in the next section, have important implications for interpretation of the PISA data for Scotland.

Summary

The above outlines key aspects of how the PISA data is collected, with a summary of this complex process provided in Figure 3. This documents how there are many channels via which the final PISA sample may become unrepresentative of the population of 15-year-olds in a given country, including school/pupil exclusions, nonresponse and important nuances that emerge via the eligibility criteria. In the following section, I discuss how these factors accumulate in a case study of the PISA 2018 data for Scotland.

Anomalies in the PISA 2018 data for Scotland

High levels of pupil exclusions

The first issue which we need to highlight with the Scottish PISA data—and, indeed, for the UK as a whole—is the comparatively high rate of pupil exclusions. This is illustrated in Figure 4, with the pupil-exclusion rate plotted along the horizontal axis and the total exclusion rate (encompassing both pupil and school level exclusions) plotted along the vertical axis. The dashed lines represent the cut-off thresholds for the maximum level of such exclusions permitted by the OECD technical standards.

There are two key points to note. First, Scotland (as well as the UK overall) narrowly failed to meet the PISA technical standards on both these exclusion criteria. Specifically, within-school exclusions totalled 3.8% of the population in Scotland (3.3% for the UK as a whole) compared to a guideline maximum of 2.5%. Likewise, Scotland's (5.4%) and the UK's (5.5%) total exclusion rate also surpassed the 5% maximum specified in the PISA technical standards (OECD, 2019: Annex I, standard 1.7). In other words, strict application of this aspect of PISA's data quality criteria would have led Scotland—and, indeed, the whole of the UK—to be removed from the study.

Second, these exclusion rates for Scotland and the UK are higher than in most other countries. Although Scotland and the UK are clearly not alone in violating the OECD's technical standards, the average within-school exclusion rate across all participating countries is substantially lower than in the UK (standing at 1.4%) as is the total exclusion rate (3.0%).

Why is this likely to be important? Such pupil-level exclusions typically occur due to issues surrounding special educational needs or recent immigrants into a country with limited language skills. These are therefore pupils who would be likely to obtain

comparatively low scores if the PISA test was accessible to them. Yet, if some countries (e.g., Scotland) are more likely to exclude such children from the sample than other countries (e.g., Japan or South Korea, where the total pupil-exclusion rate is less than 0.1%) then this is likely to introduce bias into cross-country comparisons of their PISA performance. Indeed, the OECD technical standards on exclusions are designed to limit the potential bias in the mean score from such exclusions to around five PISA test points (Rutkowski & Rutkowski, 2016). In other words, the exclusion rates observed for Scotland (and the UK as a whole) could alone lead to a non-trivial five-point decline in average PISA scores (with the standard deviation of PISA scores being approximately 100 across OECD countries).

Change of the survey date

As mentioned in the previous section, the date of the PISA study in Scotland changed in 2018. Specifically, up until 2015, PISA in Scotland was conducted between March and May, with this moving to between October and December in 2018.

Although this may seem an innocuous change at first, it potentially impacts in important ways on the Scottish PISA sample. These stem from the precise definition of the PISA target population—those aged between 15 years and three (completed) months and 16 years and two (completed) months *at the beginning of the period of testing* In other words, by altering the test dates, Scotland may have changed the way pupils being tested are grouped.

Sample design stages	Relevant technical standard
All 15-year-olds in the country	
↓	
Target population (pupils not enrolled in school excluded)	-
↓	
Target minus school exclusions	Max 2.5% of schools excluded
↓	
School sample drawn	-
↓	
School sample minus school non-respondents	Max 15% school non-response
↓	
List of all age-eligible pupils within participating schools	-
↓	
Pupil sample drawn from within participating schools	-
↓	
Pupil sample minus "ineligible" / withdrawn pupils	None
↓	
Pupil sample minus pupil-level exclusions	Max 2.5% of pupil exclusions
↓	
Pupil sample minus pupil non-respondents	Max 20% pupil non-response
↓	
Final PISA sample who take the test	

Figure 3. Illustration of key features of the PISA sample design.

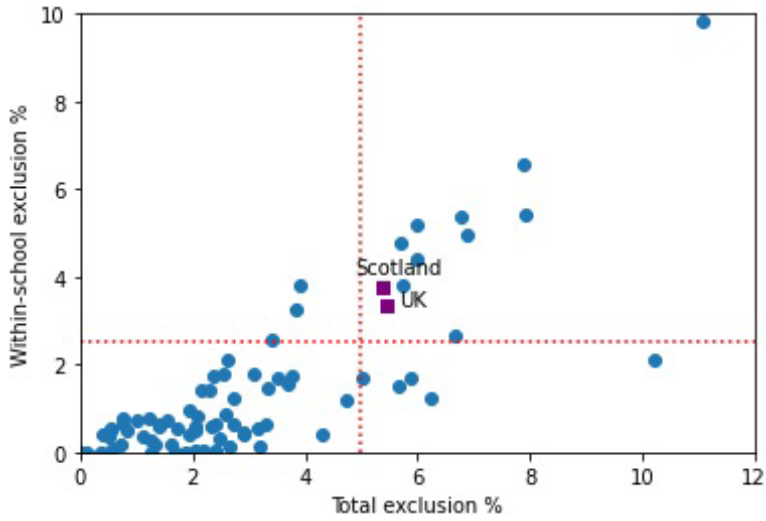

Figure 4. The within-school and total exclusion rates in PISA 2018.Notes: Dashed red lines refer to the thresholds according to the OECD technical standards. Figures for Scotland and UK as a whole are illustrated by purple square markers. Data are presented for all countries participating in PISA 2018 with data available. Source: OECD (2019: Chapter 11).

Table 1. The distribution of pupils participating in PISA across different school year groups in Scotland

Year group	2006	2009	2012	2015	2018
S3	2%	3%	3%	4%	0%
S4	89%	87%	87%	88%	50%
S5	9%	10%	10%	9%	50%

Notes: Author's calculations using the PISA 2006–2018 data for Scotland. Green shading should be read horizontally, and refers to a greater proportion of pupils belonging to that school year group.

This point is illustrated in Table 1, which presents the percentage of the Scottish PISA sample in the S4 and S5 year groups by survey round. Up until 2015, the vast majority (almost 90%) of the Scottish PISA sample were enrolled in S4, with only a small minority (just over 10%) enrolled in S5. In 2018 however, due to the change in the test date, there was an even split of the Scottish PISA sample across these two year groups (50% in S4 and 50% in S5). Of course, young people in a later school year group (S5) may well have a different distribution of academic skills than those in an earlier year group (S4). Indeed, as noted by Aloisi and Tymms (2017):

> if student results vary so much between grades, and if the proportions of students in grades also change over time, then it is reasonable to expect that fluctuations in the grade distribution might affect country outcomes.

Moreover, it is also likely that this date change—and the potential change in the composition of exactly who is being tested—may impact upon measures of

educational inequality. Unfortunately, I know of attempts by either the Scottish government or by the OECD to try and quantify the potential impact of this important change upon Scotland's PISA results.

Perhaps the most unfortunate aspect of this key change is the lack of transparency with which it (and its potential implications) have been reported. First, in the PISA 2018 report for Scotland, it is noted how the tests were conducted between October and December—but without any mention of how this was different in previous years. Second, to my knowledge, no justification has been presented as to why the test date was changed. Third, the PISA 2018 report for Scotland includes a whole section discussing issues with the interpretation of trends in PISA data over time, but completely fails to recognise this key issue.

Finally, the methodology sections of the PISA 2012, 2015 and 2018 reports for Scotland are almost identical (copied almost word for word). Importantly, the 2015 report clearly states 'students were mostly (87.5%) in the S4 year group' (Scottish Government, 2016, 10), with a similar statement in the 2012 report, that 'students were mostly in the S4 year group' (Scottish Government, 2013, 6). Yet no such statement is made in the 2018 edition. In other words, this key piece of text was selectively removed from the Scottish PISA report in 2018, in what is otherwise an almost identical passage of text. This is despite the fact that this information is clearly now more relevant than ever, given the change of test date.

High rates of pupil ineligibility/withdrawal

A further issue that may be related to the change of test month is documented in Figure 5. This plots the percentage of 'ineligible/withdrawn' pupils from PISA 2015 (vertical axis) to PISA 2018 (horizontal axis) by country. Note how Scotland is a clear

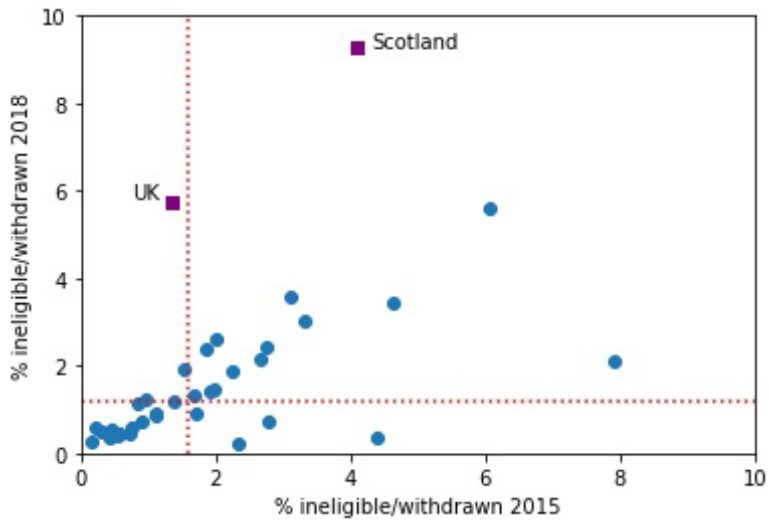

Figure 5. The percentage of pupils classified as 'ineligible/withdrawn' from the PISA 2015 and 2018 studies.Notes: Dashed red lines refer to the OECD median. Data plotted for OECD countries only. Figures for Scotland and the UK as a whole are illustrated by purple square markers. Source: OECD (2016) and OECD (2019). The Spearman correlation between data points equals 0.65.

outlier in two ways. First, the percentage of ineligible pupils in Scotland in 2018 (9.3%) is much higher than in any other country (OECD average = 1.6%; cross-country average = 1.7%). Second, the percentage of ineligible/withdrawn pupils has more than doubled in Scotland between 2015 (4.1%) and 2018 (9.3%). In comparison, in most other countries, the figures have remained broadly stable at a much lower level.

To my knowledge, this issue has not been commented upon anywhere by either the OECD or the Scottish government and there therefore seems to be no 'official' explanation as to why it has occurred. Here, I offer what I believe to be the most likely explanation.

To begin with, ineligible pupils are identified *after* the pupil sample has been drawn within participating schools. Then, according to the PISA 2018 report for Scotland, pupils may be classified as ineligible if they had left the school (between when the PISA sample was drawn and when the test was conducted):

> Students that had left the school in the interim were not considered part of the target sample. (Scottish Government, 2019, 10)

It hence seems that the high 'ineligibility' rate for Scotland in PISA 2018 is being driven by an unusually large number of pupils leaving the school between the sampling date and the PISA test window.

Why might this occur? One possibility is that this is related to when national examinations ('nationals') take place in Scotland, and young people's subsequent educational pathways. Specifically, young people in Scotland take their 'nationals' at the end of the S4 year group. Then, after completing S4, young people may decide to change schools, for instance to move to a further education college to pursue a more vocational educational pathway.

As mentioned in the sub-section above, almost 90% of the Scottish PISA sample were in S4 in PISA 2015. This means that the vast majority of pupils in Scotland had not yet taken their 'nationals' and hence were likely to still be in the same school at the time that the sampling was done and the PISA test was conducted. This changed, however, in PISA 2018 with the movement of the test date, with PISA now equally spanning S4 (pre-nationals) and S5 (post-nationals). Consequently, there may now be many more pupils in the PISA sample who have left their school after taking their nationals—thus leading to the high and rising levels of ineligibility observed in Scotland.

Importantly, those pupils who change schools between S4 and S5 are probably lower achievers; school mobility has previously been linked with lower levels of achievement (Strand & Demie, 2007), while young people who pursue vocational courses tend to have—on average—lower levels of academic achievement. In other words, Scotland's high levels of pupil ineligibility in PISA 2018 may have led to Scotland removing some lower achieving pupils from the sample.

Unfortunately, without any further detail available on what exactly is driving the high ineligibility rate in Scotland, it is difficult to say for certain why it has occurred and to fully appreciate the consequences of it. To try and find out more, I made a freedom of information request to the Scottish government—with the full list of questions asked and responses provided available from https://www.whatdotheyknow.c

om/request/720228/response/1725609/attach/3/Response%20202100141438.pdf?cookie_passthrough=1. In this, the Scottish government has confirmed how the high ineligibility figure in Scotland is 'likely to reflect the change in the timing of the PISA assessments in Scotland', with the PISA pupil lists provided during the school summer holidays and before the census at the start of the new academic year. They have therefore now confirmed the explanation that I offered above—that the high level of ineligibility has been driven by pupils moving between schools, most likely between S4 and S5. Yet they also go on to note how they are unable to precisely quantify the extent of this problem, because they 'do not hold information on how many of the ineligible students had left school between the sampling and the assessment dates'.

Low pupil response rates

As noted previously, the OECD's technical standards require that 'the final weighted student response rate is at least 80% of all sampled students across responding schools' (OECD, 2019: technical standard 1.11). An important caveat to this point, however, is that within-school exclusions and pupils deemed ineligible (as outlined in the sub-sections above) are *not* counted in these figures. Likewise, pupils in schools with low levels of participation are also not included in the official pupil response rate calculation. Thus, in reality, the technical standard applied is not 80% of all sampled pupils. Rather, it is 80% of those who were sampled, and not already excluded by their schools (due to, for instance, special educational needs), and in schools where pupil participation rates exceed 50% (explained in more detail below).

Nevertheless, Figure 6 illustrates how each country performed against this technical standard in PISA 2018. From this, there are three key points to note. First, only one (Portugal) out of the 80 participating countries failed to reach the 80% threshold. This could either be seen as a triumph of PISA in encouraging pupils to respond or, as Anders *et al.* (2021) argue, the fact that the 80% response rate threshold is too low, and not a sufficiently robust criteria to inspire confidence in the sample being representative. Second, other than Portugal, Scotland had the lowest pupil response rate of any participating country (80.5%). Finally, the pupil response rate was also low for other parts of the UK, with the figures for England (83.2%), Wales (85.5%) and Northern Ireland (83.7%) each below the OECD average (90%).

There are, however, some questions as to whether the true pupil response rate in Scotland is even lower—and that they have apparently (just) managed to reach this technical standard due to a subtle technicality in the way the pupil response rate has been calculated. In particular, the PISA 2018 report for Scotland notes:

> In total, 3767 students were deemed eligible to take part[4]

It then states:

> Of these, a total of 2969 students took part

This would therefore give an unweighted response rate of 78.8% for Scotland, falling just below the 80% threshold.[5]

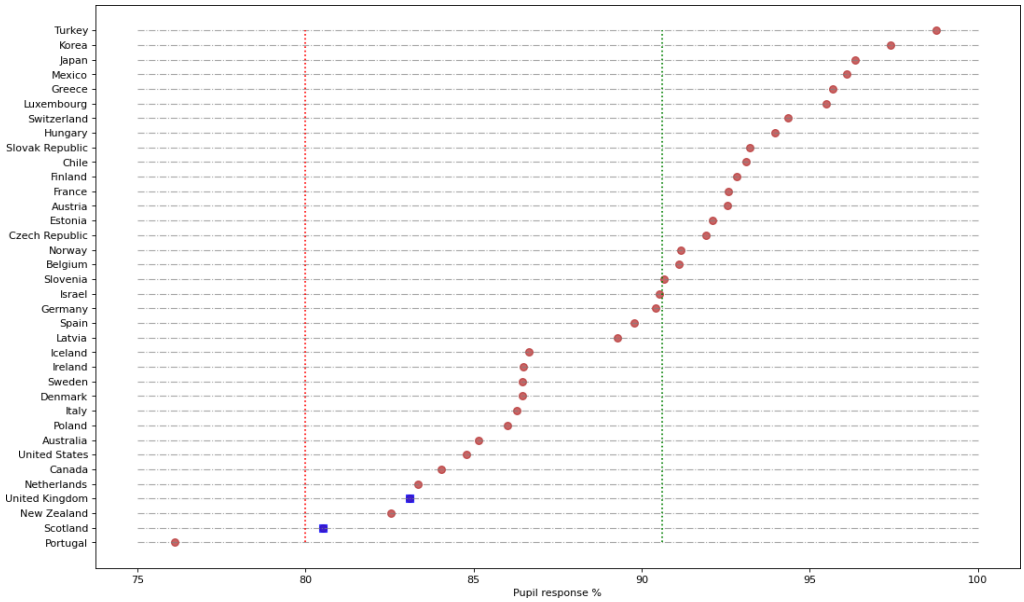

Figure 6. Pupil response rates in PISA 2018.Notes: Dashed red line refers to the minimum threshold set by the technical standard. Green line indicates the OECD median pupil response rate. Figures for the UK as a whole and Scotland are illustrated using a blue square.

How has this discrepancy occurred? Using the figures provided in the PISA 2018 report for Scotland (Scottish Government, 2019) and in response to my freedom of information request (https://www.whatdotheyknow.com/request/720228/response/1725609/attach/3/Response%20202100141438.pdf?cookie_passthrough=1), the decline between 3767 eligible participants and 2969 in the final sample seems to have occurred via:

a Pupils and/or parents opted out of the study (n = 122)
b Absence on the day of the test (n = 596)
c Eighty pupils whose status has not been accounted for.

It seems that these 80 pupils belong to schools with particularly low pupil rates.[6] In PISA, individual schools where only 25–50% of sampled pupils complete the test are incorporated into the *school* nonresponse figures—*not* the pupil response rate—despite these schools/pupils being included in the final database (and thus contributing to Scotland's PISA scores). This has led to the OECD excluding 29 (responding) and 51 (nonresponding) pupils from the calculation of the pupil response rate in Scotland, with their two schools included in the school-level nonresponse calculation instead. I discuss this issue in further detail in Appendix 2. In this I note that if these 80 pupils were included in the pupil nonresponse figures (which is arguably more appropriate), then Scotland's response rate would be 79.6%—below the 80% threshold.

It is my view that—given this accounting anomaly—a pupil-level NRBA should have been conducted by Scotland for its PISA 2018 data; it either fell very narrowly

above or narrowly below the desired 80% threshold, depending on exactly how one chooses to do the nonresponse calculation. In particular, assuming that pupil nonresponse is not random (for instance, because low-achieving and disadvantaged 15-year-olds are much more likely to be absent from school than high-achieving, socio-economically advantaged 15-year-olds), the fact that one-in-five did not complete the test has clear potential to bias the PISA results. It thus seems important that the magnitude of such potential bias is investigated and transparently reported—regardless of whether it falls just above or just below this 80% threshold. Unfortunately, to my knowledge, neither the Scottish government nor the OECD has investigated this issue, or published any such evidence in the public domain.

The overall impact of the above: low coverage of the target population

Thus far I have considered these issues in isolation. Yet what really matters is their cumulative impact. When they are taken together, how far has the PISA sample moved away from the target population?

Evidence is presented on this matter for Scotland and the UK as a whole in Table 2.[7] The first row provides an estimate of the number of 15-year-olds living in Scotland and the UK (drawn from OECD, 2019: Chapter 11). Then, moving down the rows, it provides an indication of the reduction in the target population through to the final (weighted) PISA sample, due to all the various issues discussed above (and overviewed in Figure 1). The information in Table 2 has been drawn from OECD (2019: Chapter 11) and provides, in my view, the most comprehensive picture of how the various forms of nonparticipation (exclusions, ineligibility, school nonresponse, pupil nonresponse) affect the PISA sample.

The first key point to take from Table 2 is that, for Scotland and the whole of the UK, almost 40% of the target population gets removed from the (weighted) PISA sample. If this 40% is not a random selection (and, as argued above, there is good reason to believe that they will tend to be lower-achievers) then this large amount of non-participation has clear potential to introduce bias into the results.

The second key point, importantly, helps illustrate how the main problem in Scotland—and the UK as a whole—occurs at the *pupil* level, not at the *school* level. In other words, a lot more pupils are missed out from the target population due to non-participation amongst pupils, rather than nonparticipation by schools.[8] Take the figure for Scotland, for example. A total of 5741 (53,398–47,657) pupils are missed out from the target population due to either school-level exclusions or school nonresponse. This compares to 14,147 (47,657–33,510), due to nonparticipation amongst pupils.

This is important because almost all the adjustments the OECD makes to attempt to control for selective nonparticipation occurs at the *school*-level (through the use of replacement schools and nonresponse adjustments incorporated into the weights). Almost no adjustment is made for nonparticipation at the *pupil*-level (other than for some very basic allowance of differential nonresponse by grade and gender), despite this being—as Table 2 illustrates—where most of the potential problems occur.

Table 3 helps to drive home the importance of this point for the UK. Specifically, it demonstrates how Scotland—and the UK as a whole—perform on this metric,

Table 2. The cumulative impact of exclusions, ineligibility, withdrawals and nonresponse on the PISA data for Scotland and the whole of the UK

	Whole of the UK	Scotland
Total number of 15-year-olds	703,991	53,398
Minus pupils not enrolled in school		
Total target population	697,603	53,271
Minus school-level exclusions		
Total students in all sampled schools	682,276	52,369
Minus nonresponding schools (after replacements)		
Total students in responding schools	590,558	47,657
Minus ineligible, withdrawn and within-school excluded pupils		
Total students in responding schools, less ineligible/exclusions	514,975	41,621
Minus pupil nonresponse		
Weighted final sample	427,944	33,510
Weighted final sample as a percentage of target population	61%	63%

Note: Figures refer to the estimated total population size.Source: OECD (2019: Chapter 11).

relative to all other participating countries. Scotland and the UK both fall towards the very bottom of this table; the culmination of the various forms of nonparticipation is a bigger issue here than elsewhere. Specifically, note how the final weighted PISA sample as a proportion of the target population is a lot smaller in Scotland (63%) and the UK (61%) than the international median (85%), with only a handful of nations coming out worse.

The next obvious question to ask is how much bias this problem might bring into Scotland and the UK's PISA scores? The challenge, of course, is that we do not know exactly how these nonparticipants (c. 40%) would have performed on the PISA test, had they taken it. A strong case can be made, however, that their average score would be lower than those who did take the test. For instance, in the previous sections we have discussed how the major issue in Scotland and the UK was pupil (rather than school) nonparticipation, with this driven by a combination of within-school exclusions (often due to special educational needs), absence on the day of the test and ineligibility (likely driven by pupils leaving the school). These are all characteristics that have consistently been shown to be associated with lower test scores in the education literature (Strand & Demie, 2007; Department for Education, 2015; Department for Education, 2020). Moreover, in the next section, we provide further evidence that the PISA sample has ended up underrepresenting lower-achieving pupils, at least in England and Wales.

With this assumption in hand, we follow the simulation approach outlined in Anders *et al.* (2021), who performed a similar analysis for the Canadian PISA 2015 data. The intuition behind this approach is that a value is imputed for nonparticipants under different assumptions about how they would have performed had they taken the PISA test. Specifically, the values are a random draw from a normal distribution, with the mean set to a particular (assumed) PISA score for the nonparticipants, and the standard deviation set to the value in the final sample. For simplicity, we

Table 3. Cumulative impact of exclusions, ineligibility, withdrawals and nonresponse across countries

Country	Weighted final sample as a percentage of target population	Country	Weighted final sample as a percentage of target population
Croatia	100%	Spain	84%
Macao (China)	98%	Ireland	84%
Indonesia	97%	Qatar	84%
Moldova	94%	Cyprus+	84%
Brunei Darussalam	92%	Romania	84%
Belarus	92%	Turkey+	84%
Chile	92%	Japan	84%
Montenegro	92%	Malaysia	83%
Saudi Arabia	91%	Malta	83%
Kazakhstan+	91%	Belgium	83%
Ukraine	91%	Norway+	82%
Singapore	91%	Thailand	82%
Finland	91%	Poland	81%
Morocco	90%	Italy	80%
Russia	90%	Philippines	80%
B-S-J-Z (China)	89%	Lebanon+	80%
Hungary	88%	Mexico	80%
Albania	88%	Baku (Azerbaijan)	80%
Chinese Taipei	88%	Slovak Republic	79%
Germany	88%	Iceland+	79%
Czech Republic	88%	Israel+	78%
Luxembourg+	88%	Costa Rica	75%
Greece	88%	Australia +	74%
Kosovo	88%	Sweden+	74%
North Macedonia+	88%	Colombia	74%
France	88%	Uruguay	73%
Georgia	87%	Viet Nam+	73%
United Arab Emirates	87%	Latvia	72%
Lithuania	87%	Denmark+	72%
Estonia	87%	Argentina	69%
Austria	87%	New Zealand+	68%
Peru	87%	Netherlands*+	67%
Switzerland+	86%	Hong Kong*	67%
Serbia	86%	Canada+	63%
Korea	86%	**Scotland**	**63%**
Bosnia and Herzegovina	85%	Portugal+*	62%
Dominican Republic	85%	**United Kingdom***	**61%**
Bulgaria	85%	Panama	57%

Country	Weighted final sample as a percentage of target population	Country	Weighted final sample as a percentage of target population
Jordan	85%	United States*	57%
Slovenia	84%	Brazil	56%

Notes: Green/red shading indicates smaller/bigger cumulative impact. Figure for Croatia rounded down from 107% to 100% (likely due to inaccuracies in data on total population size).
*Country had to conduct a NRBA.
+Other issues with data 'adjudicated' by the OECD (usually due to thresholds stipulated within technical standards not met).
Siginificance of bold text used to highlight Scotland and the UK.

implement this simulation using just the first plausible value. Further details about this approach can be found in Anders *et al.* (2021).

Results from this simulation—focusing upon reading—are presented in Table 4 for Scotland (Panel a) and the UK (Panel b). Column 2 documents the average PISA score assumed for the 40% of nonparticipants, with Column 1 detailing the correspondence between this and the percentile of the observed PISA reading distribution in Scotland/the UK. Column 3 provides assumptions made about the selectivity of nonparticipants. Additionally, Columns 4, 5 and 6 provide analogous results for how the scores of low-achievers (P10), high-achievers (P90) and educational inequality (P90–P10) would be affected, respectively. This, in-turn, provides some indication of the sensitivity of the UK and Scotland's PISA 2018 results to the problem of nonparticipation.

For both Scotland and the UK as a whole, average PISA reading scores could change quite dramatically under what I consider to be plausible scenarios surrounding the selectivity of nonparticipants. For instance, under the assumption that nonparticipants would have achieved an average score of 466 on the PISA reading test (i.e. around the 35th percentile of those that actually took the test), then the average PISA reading score for Scotland would drop by 13 points—from 504 down to 491. A similar decline, made under similar assumptions, would be observed for the UK as a whole.

There is, of course, a large degree of uncertainty surrounding such results, as our simulation results in Table 4 reflect. The bias brought about by such nonparticipation might be higher or lower, depending on the exact characteristics of the nonparticipants who have been selected. Yet this also clearly illustrates how the great amount of nonparticipation means that there is quite substantial uncertainty surrounding the UK's PISA results.

Anomalies in the PISA 2018 data for the rest of the UK

England

There are two specific concerns with the PISA 2018 data for England. The first, as illustrated by Figure 2, is the high level of school nonresponse. In particular, England

Table 4. Simulated PISA reading scores under differing assumptions about the likely average scores of nonparticipants

(a) Scotland

1. Nonparticipants' achievement as a percentile of observed country distribution	2. Assumed average PISA score of nonparticipants	Revised PISA scores			
		3. Mean	4. P10	5. P90	6. P90–P10
Original	—	504	383	627	244
45	492	501	377	625	248
40	480	496	371	622	251
35	466	491	365	618	253
30	452	486	358	615	257
25	439	481	351	612	261
20	424	475	342	611	269
15	405	468	329	608	279

(b) United Kingdom

1. Nonparticipants' achievement as a percentile of observed country distribution	2. Assumed average PISA score of nonparticipants	Revised PISA scores			
		3. Mean	4. P10	5. P90	6. P90–P10
Original	—	505	372	632	260
45	493	500	370	628	259
40	478	494	364	624	261
35	464	489	357	620	263
30	450	483	349	616	267
25	435	477	340	613	273
20	417	471	329	611	282
15	397	463	316	608	293

Notes: Details about the simulation available in *author cite*. Column 1 refers to the percentile of the Scottish/UK PISA reading score distribution that the average nonparticipant would have achieved had they sat the test (Column 2 illustrates the actual PISA score this corresponds to). Columns 3 to 6 then illustrate how PISA reading scores for Scotland/the UK would change under the different scenarios.

failed to meet the OECD's technical standards, and was required to conduct a school-level nonresponse bias analysis. Unfortunately, the results of this analysis have not been published by either the Department for Education or the OECD. The PISA 2018 technical report (OECD, 2019: Chapter 14) merely states:

> A school-nonresponse bias analysis was submitted, limited to England (the largest subnational entity within the United Kingdom) and relying on a direct measure of school performance in a national assessment. This analysis investigated differences between responding and nonresponding schools and between originally sampled schools and replacement schools. This supported the case that no notable bias would result from nonresponse.

With a similar nebulous statement in the PISA 2018 national report for England:

> The OECD's Technical Advisory Group was satisfied that this analysis demonstrated that no notable bias would result from the nonresponse.

This approach lacks transparency. What analysis was performed? What results were obtained? What criteria were used to reach the conclusion of there being 'no notable bias'?

I have consequently used Freedom of Information legislation to obtain a copy of the NRBA submitted by England to the OECD. Full details of the NRBA I received in response are available from https://www.whatdotheyknow.com/request/pisa_2018_data_3?nocache=incoming-1716651#incoming-1716651. Table 5 summerises what I consider to be the most important points from the NRBA that was produced, focusing upon the comparison between responding and nonresponding schools in terms of their prior achievement in national examinations. These are the school-level

Table 5. Prior achievement of responding and nonresponding schools in England

(a) Responding versus nonresponding schools

School attainment	Nonresponding	Responding (after replacement)
Top 60% of Attainment 8 distribution in 2016	38% (9)	66% (115)
Bottom 40% of Attainment 8 distribution in 2016	46% (11)	31% (55)
Missing	17% (4)	3% (5)
Total	100%	100%

(b) Comparison to originally sampled schools

School attainment	1. All originally sampled schools	2. Participating schools before replacements included	3. Participating schools after replacements included	4. Participating schools after replacements included. Weights applied
Top 60% of Attainment 8 distribution in 2016	62%	69%	66%	62%
Bottom 40% of Attainment 8 distribution in 2016	33%	28%	31%	35%
Missing	5%	4%	3%	4%
Observations	199	144	175	175

Notes: Figures in panel (a) drawn from NFER (2021a: Table 2.4). Note that I have corrected what I believe to be an error in the table submitted in the NRBA, where the figures provided in the 'responding' column did not sum to 100%. 'Nonresponding' refers to the initially drawn sample of schools that did not get replaced. Participants refers to all participating schools (whether main sample or replacement). Figures in panel (b) drawn from NFER (2021a: Table 4.4). Unweighted number of schools reported in Column 4, whereas the original NRBA presented the weighted number.

variables investigated in the NRBA that are most likely to be strongly correlated with PISA scores, and have been specifically mentioned by the OECD in their data adjudication notes for the UK (see quote above).

Panel (a) of Table 5 compares responding schools after replacement (i.e. the final sample of PISA schools) with nonresponding schools (i.e. initially sampled schools where no 'replacement' school could be found). The intuition behind this comparison is to see whether participating and nonparticipating schools are similar in terms of their historic performance in GCSE examinations. If so, then this would provide some reassurance that there has not been systematic selection of higher- or lower-achieving schools into or out of the PISA 2018 sample for England. Unfortunately, the NRBA uses only a rather coarse binary measure (whether the school is in the top 60% of the national GCSE achievement distribution versus the bottom 40%), when a more fine-grained measure would be preferable. The NRBA produced by NFER (2021a) notes that this was done 'in order to be able to apply statistical significance tests'. This, however, highlights a fundamental problem with the NRBA: far too much emphasis is placed upon statistical significance as a criterion (and any such tests are always going to be woefully underpowered, even if one accepts such significance tests to be valid in this context). Instead, the real focus should be upon the huge differences between responding and nonresponding schools. At the very least, one would have expected some sensitivity analyses to have been conducted here using a more fine-grained measure, given how this is probably the most important variable used in their analysis.

Nevertheless, it still provides clear evidence that schools with historically lower levels of GCSE performance were more likely to refuse to participate in PISA 2018. Specifically, 46% of nonresponding schools were in the bottom two quintiles of the school-level Attainment 8 distribution in 2016, compared to 31% of responding schools. Moreover, the full NRBA illustrates how this finding continues to hold the probability of nonresponse in logistic regressions modelling, where a host of other school-level variables are controlled.[9] In other words, the evidence to support claims of school nonresponse being unlikely to bias the sample is not as clear-cut as claimed by the quotes presented above from the PISA technical report and England's national report. At best, what has been reported by the NFER, Department for Education and OECD is only a partial reflection of the evidence that has been produced.

Table 5 Panel (b) presents some further evidence on this issue from the NRBA. This compares the distribution of this binary school-level achievement measure (plus a missing category) across four nested versions of the PISA sample:

1. All originally sampled schools
2. Responding schools before replacement schools are added
3. Responding schools after replacements are added
4. Responding schools after replacements are added and weights applied

A comparison of Column 1 with Column 2 reiterates the point made above; nonresponding schools had lower levels of prior GCSE performance. There is hence a greater share of schools in the top three achievement quartiles in Column 2 (participating schools from the original sample) than in Column 1 (the full original sample).

As noted in Background to how the PISA sample for the UK is drawn, PISA has two ways of trying to deal with such school nonresponse: (i) via allowing replacement schools to take the place of nonresponding schools and (ii) to include a nonresponse adjustment in the final weights. These are the results presented in Columns 3 and 4 respectively, which seemingly bring the figures much closer to those observed for the full original sample in Column 1.

At first glance, the similarity between Columns 1and 4 may seem reassuring. But does this really mean that all the potential problems surrounding school nonresponse have been resolved?

Unfortunately not. To understand why, recall from Background to how the PISA sample for the UK is drawn how replacement schools are selected, and how the PISA weights are constructed. With respect to the former, replacement schools are selected as those adjacent to the nonresponding originally sampled school on the sampling frame—which has been implicitly stratified by the *historic school performance* variable presented in Table 5b. In other words, the inclusion of replacement schools will *mechanically* improve the comparisons being made, as the variable in question helps to determine which replacement schools get selected. This is because the same variables are being used to adjust for nonresponse (through the selection of replacement schools) and then to also judge whether this nonresponse adjustment has 'worked'. It is therefore hardly surprising that no problem has been reported in the NRBA, because the replacement schools have been selected due to their similarity to the originally-sampled nonresponding schools.

A similar intuition holds for why applying the weights leads to the improvement in Table 5b; historic school performance (which is being used to judge the likely bias in the sample) has a direct role in how the weights (which are being used to adjust for the likely bias) are constructed. Once the weights are applied it is therefore unsurprising—and, in fact, *mechanical*—that the distribution of historic school performance (presented in Column 4) moves closer to the distribution for originally sampled schools (presented in Column 1). Further discussion is provided on this matter in Appendix 3. This point has actually been noted by other countries that have had to perform similar bias analyses, such as the United States, which states how such comparisons:

> may provide an overly optimistic scenario, resulting from the fact that substitution and nonresponse adjustments may correct somewhat for deficiencies in the characteristics examined, but there is no guarantee that they are equally as effective for other characteristics and, in particular, for student achievement. (National Centre for Education Statistics, 2019)

If those responsible (the NFER and Department for Education) really wanted to know about the bias school nonresponse brought into the PISA sample, they would have conducted a different analysis. Table 5b would still have been produced, but using *pupil-level data* from the schools for the cohort in question (i.e. pupils in these schools who took their GCSE in 2019), focusing upon the distribution of Key Stage 2 scores and/or their final GCSE grades.[10] This approach would have two key advantages. First, by using pupil (rather than school) level data, the analysis would have much more power to detect potential differences. Second, it would illustrate potential

bias in a key variable that has not been directly used in the selection of replacement schools or in the construction of the response weights. It would not suffer the problem of the same variable (school-level historic GCSE performance) being used to both adjust for nonresponse and then also to judge whether that nonresponse adjustment has 'worked'.

My interpretation of the available evidence on potential bias in the PISA sample for England from school nonresponse is hence not as optimistic as the views of the OECD or as those presented in England's national report. Of course, such matters are never black or white, and are often a matter of judgement and opinion about the evidence available. Yet this helps to iterate a recurring theme presented throughout this paper. In order for academics and policymakers to come to their own reasoned judgements on such issues, it is vital that the evidence is openly and transparently reported when the PISA results are released, as a matter of course. Unfortunately this is not currently the case, with little more than a nebulous paragraph about such issues relegated to the annexes of the reports—with no hard data presented to support the claims being made.

The second major issue for the PISA 2018 data for England can be inferred from Tables 2 to 4. Specifically, as England dominates the UK figures (making up 84% of the weighted sample), it becomes clear that there has been significant nonparticipation in the study in England. This has occurred through various channels, and is *not* primarily driven by the issue of school nonresponse, as discussed above. In particular, as can be inferred from Table 2, England not only had high levels of school nonresponse, but also high-levels of within-school pupil exclusions and pupil nonresponse. Thus, as can be inferred from Tables 3 and 4, the PISA data for England suffers the same challenges as the data for Scotland, with the various forms of nonparticipation having a large cumulative impact upon the sample (Table 3), which means that there is quite a high degree of uncertainty over England's PISA scores (Table 4).

Importantly, however, it is possible to investigate potential bias in the PISA sample for England in one additional way. As part of my freedom of information request, I additionally asked for the GCSE grades obtained by the PISA 2018 cohort (examinations which had taken place just six months after they took the PISA test).[11] These can then be compared to the national distribution of GCSE grades for 16-year-olds which, unlike PISA, is based upon data from *all* Year 11 pupils (and thus do not suffer from issues such as school nonresponse, pupil exclusions or pupil nonresponse). The results of this comparison can be found in Table 6, Panel (a).

The PISA 2018 data for England clearly underrepresents lower-achieving pupils, compared to the official GCSE mathematics grade distribution. In total, 21.3% of the PISA 2018 sample for England failed to achieve a Grade 4 in their GCSEs. This compares to 28.5% of 16-year-olds across England as a whole in 2019. Moreover, there are particularly pronounced differences at the lowest grade levels, including at Grade 1 (3.2% in PISA *versus* 5.4% in the national distribution), Grade 2 (6.7% *versus* 8.6%) and Grade 3 (11.3% *versus* 12.7%). The opposite then holds true at the higher grade levels, with the PISA 2018 cohort having notably more pupils at Grade 5 and above than can be observed in population-level data.

How much more of an impact is this bias likely to have on average PISA scores in England? To address this issue, the final column of Table 5a presents average PISA

Table 6. GCSE mathematics grade distribution. Comparison of the PISA 2018 sample to the national grade distribution

(a) England

Mathematics	PISA	Official	Difference	Average PISA score by grade
9	4.0%	3.7%	0.3%	620
8	8.4%	7.2%	1.2%	585
7	10.9%	9.5%	1.4%	561
6	12.7%	11.5%	1.2%	532
5	20.2%	18.2%	2.0%	506
4	22.5%	21.4%	1.1%	467
3	11.3%	12.7%	−1.4%	434
2	6.7%	8.6%	−1.9%	396
1	3.2%	5.4%	−2.2%	364
Ungraded	0.1%	1.8%	−1.7%	337
	100.0%	100.0%		

(b) Wales

Mathematics	PISA	Official data	Difference	Average PISA score by grade
A*	11.1%	8.7%	2.4%	577
A	12.0%	9.2%	2.8%	541
B	19.4%	17.1%	2.3%	504
C	26.2%	22.9%	3.3%	474
D	13.5%	13.9%	−0.4%	437
E	7.9%	10.4%	−2.5%	413
F	3.7%	5.9%	−2.2%	387
G	3.3%	5.9%	−2.6%	382
Ungraded	2.9%	6.1%	−3.2%	359
	100.0%	100%		

Notes: Green/red shading illustrates where the percentage achieving the grade is higher/lower in PISA than the national grade distribution. Figures for England based upon https://schoolsweek.co.uk/gcse-results-2019-mathematics/ and https://www.whatdotheyknow.com/request/pisa_2018_data_3?nocache=incoming-1716651#incoming-1716651. Figures for Wales refer to GCSE mathematics alone (not numeracy). For Wales, data on PISA grade distribution based on freedom of information request submitted by the author https://www.whatdotheyknow.com/request/pisa_2018_data_2. Data on 'official' grade distribution taken from https://statswales.gov.wales/Catalogue/Education-and-Skills/Schools-and-Teachers/Examinations-and-Assessments/Key-Stage-4/gcsentriesandresultspupilsaged15only-by-subjectgroup, using data for the 2018/19 academic year. Average PISA scores by grade based upon Gambhir *et al.* (2020: Table 3.3) and refers to data on best GCSE grade out of numeracy and mathematics.

scores by GCSE grade, based upon data gathered as part of another freedom of information request (https://www.whatdotheyknow.com/request/average_pisa_scores_by_gcse_grad?nocache=incoming-1736259#incoming-1736259). I take these figures and create two weighted average PISA scores, based upon the GCSE grade distribution in (i) the PISA data and (ii) official population data. The difference between these two weighted averages is then used to estimate the magnitude of the upward bias in average PISA mathematics scores. Overall, I estimate that average PISA

mathematics scores are around 11 points too high in England, compared to a truly representative sample from the population having participated. Unfortunately, it is not possible to establish what is driving this bias—whether it is due to school nonresponse, exclusions of pupils by schools from the study or pupil absence/refusal to take part.

Wales

At first glance, the PISA 2018 data for Wales may seem to compare reasonably well to other parts of the UK. Unlike England and Northern Ireland, the school response rate met the OECD's technical standards (though only after replacement schools are included). The pupil response rate in Wales (85.5%) was also higher than in Northern Ireland (83.7%), England (83.2%) and Scotland (80.5%), although still reasonably low by international standards. However, as Wales is not a full participant in PISA (i.e. it is not an 'adjudicated sub-region'), little is known about the proportion of excluded or ineligible pupils.

As the PISA 2018 data for Wales has also been linked to pupils' administrative records, it is possible to compare the GCSE grades they achieved to the national grade distribution. This, in turn, can be used to provide some insight into whether bias may have crept into the Welsh PISA sample. The results from this analysis can be found in the Table 5 Panel (b).

A key finding from this table is that the PISA sample for Wales seems to have a disproportionate share of high achievers, while systematically underrepresenting those who achieve low GCSE grades. For instance, according to official government data, 41% of young people in Wales failed to achieve a GCSE C grade in mathematics in the 2018/19 academic year. Yet, the figure for the PISA 2018 cohort was just 31%. From Table 5b, one can see that PISA particularly underrepresents those who were ungraded (2.9% compared to official figures of 6.1%), Grade G (5.9% *versu*s 3.3%) and Grade E (7.9% *versus* 10.4%). On the other hand, there is clear evidence of inflation in PISA at Grade C (26.2% *versus* official figures of 22.9%) and Grade A (12% *versus* 9.2%). Analogous figures have been obtained for other subjects (e.g., GCSE numeracy and English language) and produced similar results. Table 5b therefore provides unequivocal evidence of bias in the PISA sample for Wales.

What impact is this likely to have had on average PISA scores in Wales? To address this issue, I draw upon data from Gambhir *et al.* (2020: Table 3.3)—which provides information on average PISA scores according to GCSE grades in Wales—to apply a similar method of estimating the bias in average PISA scores as outlined for England in the sub-section above. Overall, I estimate that average PISA mathematics scores are around 14 points too high in Wales, compared to a truly representative sample from the population having been drawn. This is substantial, given that the standard deviation of PISA scores across OECD countries is approximately 100 points, and further illustrates how caution is needed when interpreting the PISA 2018 data for Wales.

Northern Ireland

As illustrated by Figure 2, school nonresponse was significantly higher in Northern Ireland than the rest of the UK. In fact, if just one more originally sampled school had refused to take part, Northern Ireland's before replacement response rate would have fallen to below 65%, which would have been considered 'not acceptable' (if judged against the OECD's technical standards).[12] Moreover, the use of replacement schools (and the nonresponse adjustment incorporated into the PISA weights) is likely to be a less successful strategy in guarding against nonresponse bias in Northern Ireland than in England, Wales and Scotland. This is because the stratification variables used in Northern Ireland—which play a key role in PISA's nonresponse adjustments—do *not* include any information on historical school performance in GCSE examinations (or equivalent), unlike the rest of the UK (see Appendix 1 for details).

If Northern Ireland were an adjudicated sub-region in PISA, the OECD would have required a school-level NRBA to take place. However, as Northern Ireland does not participate in PISA as an independent nation, this was not required by the OECD.

Yet the PISA 2018 report for Northern Ireland clearly states that such a NRBA did take place (Sizmur *et al.*, 2019c):

> The OECD required a NRBA for England because England represents 84% of the UK weighted sample (Scotland 8%; Wales 5%; Northern Ireland 3%). As the response rate for NI was also below the OECD's requirements, a further NRBA was carried out for NI, although not required by OECD.

And it goes on to state how:

> The results of both NRBAs were positive, meaning that the samples for UK and NI were representative and not biased.

However, the actual results from this NRBA were not provided in Northern Ireland's PISA report. This therefore leaves open many questions—how was this analysis conducted and by whom? What does 'positive results' mean? Who exactly reviewed this document and how was this judgement reached?

It also conflicts somewhat with the OECD's technical report on the NRBA that was submitted for the UK, noting that the evidence that they made a judgement on was limited to England only (OECD, 2019: Chapter 14):

> A school-NRBA was submitted, limited to England (the largest subnational entity within the United Kingdom) and relying on a direct measure of school performance in a national assessment.

There are therefore again some issues surrounding transparency of reporting. In order to find out more, I submitted a Freedom of Information request to seek clarification on these key points (https://www.whatdotheyknow.com/request/pisa_2018_data#incoming-1699808). The document received in response can be found at https://www.whatdotheyknow.com/request/pisa_2018_data?nocache=incoming-1717380#incoming-1717380. This confirms that:

- The NRBA for Northern Ireland was *not* sent to the OECD.

- The NRBA was undertaken by the National Foundation for Educational Research (NFER), who were the contractors for the PISA 2018 study and was shared just with UK government officials.
- It was therefore a combination of UK government officials and the NFER who judged the results of the NRBA, showing 'positive' results (though the split of responsibilities remains unclear).
- Critically, there was no outside scrutiny of the NRBA produced (not even by the OECD).

What about the evidence presented in the NRBA itself? Was it really as 'positive' as claimed in the national report?

The NRBA for Northern Ireland followed the approach that was used for England, as described above. Participating schools were compared to nonparticipating schools to see if they were similar in terms of observable characteristics. The distribution of these characteristics were then compared across the original sample and participating schools (both before and after replacement schools were included, and with and without weights applied). I summarise what I believe to be the key figures from this NRBA in Table 7.[13]

There are two key points to note. First, very few characteristics of schools have been compared. In particular, the only variables considered are gender (boys-only school, girls-only school, mixed), region and school-type. The clearest—and most important—difference to the NRBA conducted for England is that no information on historic school performance in GCSE examinations is included. Second, the sample size is small—only 102 schools at most—with what seems to be substantial reliance upon whether differences are 'statistically significant' or not. There are of course questions about whether such significance tests are even valid in such a context (Gorard, 2010). Yet even if one accepts significance tests are a valid approach here, with only around 100 observations, any such tests will be very underpowered. In other words, this combination of an investigation of limited characteristics and reliance upon statistical significance means it is almost impossible to detect if any bias is present or not.

It therefore seems that for Northern Ireland an absence of evidence is being used to claim that there is absence of bias. The problem is that the investigations of potential bias have been extremely limited, with an almost impossibly high bar set. Again, as with the bias analysis conducted for England, the actual results of the analysis are open to interpretation, with different individuals likely to form different opinions based upon the evidence available. Yet, as I argued above in the case of England, it is critical that such evidence is clearly and transparently reported, so that independent judgements can be formed. Relegating such information to a couple of nondescript sentences in the appendix—simply saying that the results are 'positive' and that the sample is unequivocally 'representative'—should not be considered acceptable.

Note that an additional issue with the Northern Ireland data is that no information is published about the number/proportion of school exclusions, within-school exclusions or ineligible pupils. It is therefore not possible to provide comparable figures to those for Scotland and the UK presented in Tables 2–4. And it is therefore, overall,

Table 7. Prior achievement of responding and nonresponding schools in Northern Ireland

(a) Responding versus nonresponding schools

	Nonresponding	Responding (after replacement)
School type		
Grammar	30%	43%
Non-grammar	70%	57%
Region		
Belfast	Suppressed	Suppressed
North Eastern	39%	17%
South Eastern	22%	17%
Southern	22%	27%
Western	Suppressed	Suppressed
School gender		
Female	0%	17%
Male	9%	14%
Mixed	91%	70%
N	23	79

(b) Comparison to originally sampled schools

	1. All originally sampled schools	2. Participating schools before replacements included	3. Participating schools after replacements included	4. Participating schools after replacements included. Weights applied
School type				
Grammar	40%	46%	43%	34%
Non-grammar	60%	55%	57%	67%
Region				
Belfast	19%	24%	20%	17%
North Eastern	22%	14%	17%	17%
South Eastern	18%	14%	17%	18%
Southern	26%	29%	27%	30%
Western	17%	20%	20%	18%
School gender				
Female	13%	20%	17%	15%
Male	13%	17%	14%	15%
Mixed	75%	64%	70%	71%
N	102	66	79	79

Notes: Figures refer to column percentages. Figures may not sum to 100% due to rounding. Small figures have been supressed in places. Figures in panel (a) drawn from NFER (2021b: Tables 2.1–2.3). 'Nonresponding' refers to the initially drawn sample of schools that did not get replaced. Participants refers to all participating schools (whether main sample or replacement). Figures in panel (b) drawn from NFER (2021b: Tables 4.1–4.3). Unweighted number of schools reported in Column 4, whereas the original NRBA presented the weighted number.

difficult to estimate the cumulative impact that the various forms of nonparticipation has had upon Northern Ireland's PISA results.

Conclusions

PISA is a widely watched study of 15-year-olds' abilities to use their reading, mathematics and science knowledge and skills to meet real-life challenges. Since its inception in 2000, it has had a major impact upon governments and education policy, driving changes to schooling systems across the globe. In the United Kingdom, PISA has become the main resource to compare inputs and outcomes across its four devolved nations, representing the only UK-wide assessment taken by a sample of pupils on a regular basis. The triennial PISA results have thereby become a high-profile issue in all four nations—England, Northern Ireland, Scotland and Wales.

Unfortunately, PISA has a rather chequered history in the UK. Specifically, after the UK was excluded from the results of the 2003 edition due to concerns over low response-rates and data quality, the validity of the PISA study was brought into question (Jerrim, 2013). Many assume that such issues are now in the past, given how the data for the UK has always been deemed to be of acceptable quality in all subsequent PISA cycles. Yet, in reality, the situation is much more complex than first meets the eye. There remains many ways for countries to not test pupils who are technically part of the target population, with lower achievers disproportionately likely to be removed from the sample. The aim of this paper has been to explain how such issues arise based upon a case study of the PISA 2018 data for the UK. In doing so, it is hoped that the paper helps to broaden understanding of these technical but important points amongst a wider audience.

The paper illustrates how the UK—and, by implication, its four constituent nations—have some of the lowest overall participation rates of any country. Importantly, this nonparticipation seems to be mostly driven by the selection of pupils from the sample (e.g., pupils not turning up on the day of the PISA test) rather than by schools. This occurs through various channels, including schools excluding certain pupils from the test, pupils being classified as ineligible due to school moves and non-response. Moreover, for some parts of the UK, there is clear evidence that this has a non-trivial impact upon the PISA data being representative, leading to a sizable upward bias in average scores. For instance, I estimate that average PISA mathematics scores in Wales would likely be around 15 points lower if a truly representative sample of pupils had taken the test.

The paper has also raised some issues surrounding transparency of reporting the PISA results. In Scotland, important changes were made to PISA in 2018—such as changing when the test is taken. Yet this change, and its implications, have not to my knowledge been documented by either the Scottish government or by the OECD. Likewise, other clear anomalies with the Scottish data (e.g., the very high number of 'ineligible' pupils) have not been explained or discussed. In England, a NRBA was produced, but not published (I could only obtain it via a freedom of information request). A similar NRBA was produced for Northern Ireland, but with even less clarity about what exactly was produced and how this evidence was judged. Again, this information was only obtained via a freedom of information request. Finally, in Wales

and Northern Ireland, key pieces of information are not reported as a matter of course, such as the number of within-school exclusions and ineligibility rates. This means that we do not currently have any handle on the overall nonparticipation rates in PISA in these parts of the UK. These are all basic facts about the data that have not been transparently reported, clearly thought through or discussed.

There are of course some limitations of this work. First, although I have illustrated how nonparticipation is high across the UK—and that this clearly leads to bias in the data for at least some of the constituent nations—it has not been possible to investigate what causes it. For instance, it is not possible with the data available to establish whether it is school nonresponse, pupil nonresponse, within-school exclusions or pupil mobility that is driving the bias clearly observed in the PISA data for Wales. Further data, tracking each pupil via administrative records through each stage of the selection process outlined in Figure 2, would be needed to provide further insight into this issue. Second, the focus of this paper has been PISA 2018, and not how these issues may have affected previous PISA rounds. As this paper has illustrated, gaining access to and understanding all the nuances for even a single round of PISA is challenging. This task then gets multiplied if one attempts to consider multiple PISA sweeps. Yet building up a clearer picture on this matter is obviously important to help academics and policymakers build a better picture of the reliability of PISA to inform about changes over time. Finally, the paper has presented a case study for the UK. Such issues may of course impact upon other countries as well, particularly those with low overall participation rates, joining the UK towards the bottom of Table 3 (e.g., Canada, Sweden, New Zealand, Portugal, Hong Kong). Although not all will have access to the same high-quality national data to conduct such investigations (e.g., I have only been able to illustrate the substantial bias in the English and Welsh PISA 2018 data due to the link that has been made with GCSE grades), there nevertheless remains much value in similar research being conducted in other countries where this is possible.

Despite these limitations, the work has clear implications for policy and practice. The most pressing issue is for the UK Statistics Authority to conduct an independent review of the UK's PISA data. This should include a focus on the transparency of reporting and documentation of key issues, providing some 'best practice' guidelines for each of the four UK governments to follow in the reporting of future PISA rounds. To help facilitate this, Wales and Northern Ireland should follow Scotland's lead and apply to become 'adjudicated sub-regions' in PISA. Although this paper has illustrated how this is no panacea to all potential problems, it would ensure that some key information about the Welsh and Northern Irish data gets to be reported by the OECD, and that the PISA data for these countries will be held to the same technical standards as Scotland's and (essentially) England's. In addition, as each of the four UK nations conduct national examinations not long before/after PISA is conducted, they each have access to high-quality data to investigate and document potential bias in the sample (similar to the comparisons I have presented in Table 5). The UK thus actually has very good data to thoroughly investigate the issues discussed throughout this paper, but currently does not do so. Yet such analyses are informative, quick and simple to conduct, and should be reported for future rounds of PISA as a matter of course.

Finally, at an international level, the OECD needs to reconsider its technical standards, the stringency to apply these, and its data adjudication processes. The evidence presented in this paper illustrates how the processes currently in place flatter to deceive and are nowhere near robust enough to support the OECD's claims that PISA provides truly representative and cross-nationally comparable data.

Funding

No funding was received for this paper.

Conflict of interest

None.

Ethics approval

The research has been conducted under BERA ethical guidelines.

Data availability statement

Any data analysed in this paper is available from http://www.oecd.org/pisa/data/

NOTES

[1] The PISA technical standards note how up to 0.5% of a country's schools can be excluded due to geographical inaccessibility, extremely small schools—where administration of PISA would be not feasible—with a further 2% where schools only contain students that would be classed as 'within-school exclusions' (see below for further details).

[2] Historic school achievement in GCSEs is an implicit stratification variable in England and Wales, and an explicit stratification variable in Scotland. Historic school GCSE performance is not used to stratify the sample in Northern Ireland.

[3] In PISA 2018, the testing spilled into January 2019—likely due to the problems surrounding low school response rates discussed above.

[4] Note that this figure was recorded after within-school exclusions and after ineligible pupils have been removed.

[5] The weighted response rate for Scotland reported by the OECD was 80.53%, compared to an unweighted figure of 80.51%. This illustrates how the difference between the use of weighted and unweighted figures for this purpose is—at least in the case of Scotland—trivial.

[6] This goes against what has been stated in response to Question 6 within my freedom of information request. Specifically, the Scottish government has stated: 'of the 4265 students who were originally sampled, 356 were deemed to be ineligible and legitimate nonparticipants. A further 142 students were withdrawn by the school because of their additional support needs'—thus seemingly ruling out the possibility that these 80 pupils are in any way related to school non-response.

[7] It not possible to produce analogous figures for England, Northern Ireland and Wales as individual nations as they are not official 'adjudicated sub-regions'. Thus comparable figures are not reported for these three nations by the OECD.

[8] Note the figures are reported after school replacements are included. In essence, there is an implicit assumption here that these substitute schools provide an unbiased replacement for nonresponding, initially sampled schools. Given how England, Wales and Northern Ireland all include a measure of school-performance in the stratification variables that play a key role in how the replacement schools are selected, this does not seem an unreasonable assumption to make.

[9] This includes a continuous school-level Attainment 8 measure, as well as this binary variable. Interestingly, this continuous Attainment 8 measure is also independently associated with the probability of response (statistically significant at the 10% level). This further suggests that the binary measure of historic school performance does not fully capture the issue.

[10] Although GCSE grades would be preferable—given that these exams are taken shortly after the PISA tests—this information would not have been available at the time when the NRBA was conducted. However, data on pupils' Key Stage 2 scores would have been available.
[11] The Department for Education noted that the final PISA weights were applied when producing the PISA 2018 GCSE grade distribution.
[12] Also, the unweighted response rate was 64.7%—below the 65% threshold.
[13] The full NRBA is available from https://www.whatdotheyknow.com/request/pisa_2018_data?nocache=incoming-1717380#incoming-1717380).
[14] This is inconsistent with the Scottish government's national report, which specifically stated: 'In total, 3767 students were deemed eligible participants. Of these, a total of 2969 students took part, *with the balance being those who did not wish to take part (both students and their parents were given the opportunity to opt out of the survey), those who were absent on the day of the test or were withdrawn by the school because of their additional support needs*'. There was no mention of these 80 pupils being removed from the response rate calculation, due to the low participation within two schools.
[15] Although this is an unweighted figure, the weighting would seem to make a trivial difference here. According to the OECD (2019: Table 11.8) the difference between Scotland's weighted and unweighted pupil response rate is tiny—0.02%.

References

Aloisi, C. & Tymms, P. (2017) PISA trends, social changes, and education reforms, *Educational Research and Evaluation*, 23(5–6), 180–220.

Anders, J., Has, S., Jerrim, J., Shure, N. & Zieger, L. (2021) Is Canada really an education superpower? The impact of non-participation on results from PISA 2015, *Educuational Asseessment Evaluation and Accountability*, 33(1), 229–249.

Baird, J.-A., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T. L. & Daugherty, R. (2011) *Policy effects of PISA* (Pearson UK). Available online at: https://research-information.bris.ac.uk/en/publications/policy-effects-of-pisa (accessed 2 February 2021).

Breakspear, S. (2012) *The policy impact of PISA: an exploration of the normative effects of international benchmarking in school system performance*. OECD Education Working Papers No. 71. Available online at: https://www.oecd-ilibrary.org/docserver/5k9fdfqffr28-en.pdf?expires=1612262413&id=id&accname=guest&checksum=1619E79B31CAE989D091D3BDE186E630 (accessed 2 February 2021).

Coughlan, S. (2019) *Pisa tests: UK rises in international school rankings*. Available online at: https://www.bbc.co.uk/news/education-50563833 (accessed 2 February 2021).

Department for Education (2015) *The link between absence and attainment at KS2 and KS4 2012/13 academic year*. Department for Education Research Report. Available online at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/412638/The_link_between_absence_and_attainment_at_KS2_and_KS4.pdf (accessed 28 January 2021).

Department for Education (2020) *Special educational needs and disability: an analysis and summary of data sources*. Department for Education Research Report. Available online at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/882802/Special_educational_needs_and_disability_-_an_analysis_and_summary_of_data_sources.pdf (accessed 28 January 2021).

El Masri, Y.H., Baird, J. & Graesser, A. (2016) Language effects in international testing: the case of PISA 2006 science items, *Assessment in Education: Principles, Policy & Practice*, 23(4), 427–455.

Freitas, P., Nunes, L.C., Balcão Reis, A., Seabra, C. & Ferro, A. (2016) Correcting for sample problems in PISA and the improvement in Portuguese students' performance, *Assessment in Education: Principles, Policy & Practice*, 23(4), 456–472.

Gambhir, G., Dirie, A. & Sizmur, J. (2020) *PISA 2018 additional analyses: regional performance and PISA/GCSE matching in Wales*. Available online at: https://www.nfer.ac.uk/media/4193/pisa_2018_additional_analyses_regional_performance_and_pisa_gcse_matching_in_wales.pdf (accessed 31 January 2021).

Goldstein, H. (2017) Measurement and evaluation issues with PISA, in: L. Volante (Ed) *The PISA effect on global educational governance* (New York, NY, Routledge), 49–58.

Gorard, S. (2010) All evidence is equal: the flaw in statistical reasoning, *Oxford Review of Education*, 36(1), 63–77.

Jerrim, J. (2013) The reliability of trends over time in international education test scores: is the performance of England's secondary school pupils really in relative decline?, *Journal of Social Policy*, 42(2), 259–279.

Jerrim, J. (2016) PISA 2012: how do results for the paper and computer tests compare?, *Assessment in Education: Principles, Policy & Practice*, 23(4), 495–518.

Jerrim, J., Micklewright, J., Heine, J.H., Salzer, C. & McKeown, C. (2018) PISA 2015: how big is the 'mode effect' and what has been done about it?, *Oxford Review of Education*, 44(4), 476–493.

Kankaraš, M. & Moors, G. (2014) Analysis of cross-cultural comparability of PISA 2009 scores, *Journal of Cross-Cultural Psychology*, 45(3), 381–399.

Lockheed, M., Prokic-Bruer, T. & Shadrova, A. (2015) How have PISA results informed education policy discussions and affected policy in middle-income countries?. *The Experience of Middle Income Countries Participating in PISA 2000-2015* (Washington, D.C., The World Bank/Paris, OECD Publishing), 63–75.

Machin, S., McNally, S. & Wyness, G. (2013) Educational attainment across the UK nations: performance, inequality and evidence, *Educational Research*, 55(2), 139–164.

McKeown, C., Denner, S., McAteer, S., Shiel, G. & O'Keeffe, L. (2019) *Learning for the future: the performance of 15-Year-olds in Ireland on reading literacy, science and mathematics in PISA 2018* (Dublin, Educational Research Centre). Available online at: https://www.erc.ie/wp-content/uploads/2020/07/B23321-PISA-2018-National-Report-for-Ireland-Full-Report-Web-4.pdf (accessed 5 February 2021).

Micklewright, J., Schnepf, S. & Skinner, C. (2012) Non-response biases in surveys of school children: the case of the English PISA samples, *Journal of the Royal Statistical Society. Series A (General)*, 175, 915–938.

National Centre for Education Statistics (2019) *US non-response bias analysis*. Available online at: https://nces.ed.gov/surveys/pisa/2018technotes-12.asp (accessed 7 February 2021).

National Foundation for Educational Research (2021a) *PISA 2018 bias analysis: England*. Available online at: https://www.whatdotheyknow.com/request/pisa_2018_data_3?nocache=incoming-1716651#incoming-1716651 (accessed 3 February 2021).

National Foundation for Educational Research (2021b) *PISA 2018 bias analysis: Northern Ireland*. Available online at: https://www.whatdotheyknow.com/request/pisa_2018_data?nocache=incoming-1717380#incoming-1717380 (accessed 4 February 2021).

OECD (2014) *Improving schools in Wales: an OECD perspective*. Available online at: http://www.oecd.org/education/Improving-schools-in-Wales.pdf (accessed 2 February 2021).

OECD (2016) *PISA 2015 technical report*. Available online at: https://www.oecd.org/pisa/data/2015-technical-report/ (accessed 27 January 2021).

OECD (2019) *PISA 2018 technical report*. Available online at: https://www.oecd.org/pisa/data/pisa2018technicalreport/ (Accessed 27 January 2021).

Pereira, M. (2011) *An analysis of Portuguese students' performance in the OECD programme for international student assessment (PISA)*. Available online at: https://www.bportugal.pt/sites/default/files/anexos/papers/ab201111_e.pdf (accessed 8 April 2019).

Rutkowski, L. & Rutkowski, D. (2016) A call for a more measured approach to reporting and interpreting PISA results, *Educational Researcher*, 45(4), 252–257.

Scottish Government (2013) *Programme for International Student Assessment (PISA) 2012: highlights from Scotland's results*. Available online at: https://dera.ioe.ac.uk/18906/1/00439464.pdf (accessed 27 January 2021).

Scottish Government (2016) *Programme for International Student Assessment (PISA) 2015: highlights from Scotland's results*. Available online at: https://www.gov.scot/binaries/content/documents/govscot/publications/research-and-analysis/2016/12/programme-international-student-assessment-pisa-2015-highlights-scotlands-results/documents/00511095-pdf/00511095-pdf/govscot%3Adocument/00511095.pdf?forceDownload=true (accessed 27 January 2021).

Scottish Government (2019) *Programme for International Student Assessment (PISA) 2018: highlights from Scotland's results*. Available online at: https://www.gov.scot/binaries/content/documents/govscot/publications/statistics/2019/12/programme-international-student-assessment-pisa-2018-highlights-scotlands-results/documents/programme-international-student-assessment-pisa-2018-highlights-scotlands-results/programme-international-student-assessment-pisa-2018-highlights-scotlands-results/govscot%3Adocument/programme-international-student-assessment-pisa-2018-highlights-scotlands-results.pdf?forceDownload=true (Accessed 27 January 2021).

Sizmur, J., Ager, R., Bradshaw, J., Classick, R., Galvis, M., Packer, J., Thomas, D. & Wheater, R. (2019a) *Achievement of 15- year-olds in England: PISA 2018 results*. Department for Education Research Report. Available online at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/904420/PISA_2018_England_national_report_accessible.pdf (Accessed 27 January 2021).

Sizmur, J., Ager, R., Bradshaw, J., Classick, R., Galvis, M., Packer, J., Thomas, D. & Wheater, R. (2019b) *Achievement of 15- year-olds in Wales: PISA 2018 results*. Welsh government report. Available online at: https://gov.wales/sites/default/files/statistics-and-research/2019-12/achievement-15-year-olds-program-international-student-assessment-pisa-national-report-2018_0.pdf (accessed 27 February 2021).

Sizmur, J., Ager, R., Bradshaw, J., Classick, R., Galvis, M., Packer, J., Thomas, D. & Wheater, R. (2019c) *Achievement of 15- year-olds in Northern Ireland: PISA 2018 results*. Department for Education Northern Ireland report. Available online at: https://www.education-ni.gov.uk/sites/default/files/publications/education/Achievement%20of%2015-year-old%20pupils%20in%20Northern%20Ireland%20PISA%202018%20National%20Report.PDF (accessed 27 January 2021).

Spaull, N. (2019) Who makes it into PISA? Understanding the impact of PISA sample eligibility using Turkey as a case study (PISA 2003–PISA 2012), *Assessment in Education: Principles, Policy & Practice*, 26(4), 397–421.

Strand, S. & Demie, F. (2007) Pupil mobility, attainment and progress in secondary school, *Educational Studies*, 33(3), 313–331.

Zieger, L., Jerrim, J., Anders, J. & Shure, N. (2020) *Conditioning: how background variables can influence PISA scores*. UCL CEPEO working paper 20-09. Available online at: https://repec-cepeo.ucl.ac.uk/cepeow/cepeowp20-09.pdf (accessed 4 February 2021).

Appendix 1. The explicit and implicit stratification variables used in PISA 2018 for the four UK countries

	England	Wales	NI	Scotland
Explicit	School type	School type	School type	Funding type
	Region	Region	Region	School attainment
Implicit	Gender	Gender	Gender	School gender
	School performance	School performance		Area type
	Local authority	Local authority		

Appendix 2. Differences between figures reported by the Scottish government and the OECD, with respect to the PISA 2018 sample.

Further details on the sample composition for Scotland (Scottish Government, 2019:

Chapter 1:Paragraph 13) the following figures are provided for (a) the total number of pupils that were sampled and (b) the total number of eligible participants:

- 4265 pupils drawn in the sample
- 3767 pupils deemed eligible participants

These figures have then been repeated by the Scottish government in response to my freedom of information request https://www.whatdotheyknow.com/request/720228/response/1725609/attach/3/Response%20202100141438.pdf?cookie_passthrough=1. Therefore, 498 (4265–3767) pupils have been removed after the PISA sample for Scotland has been selected (and do not contribute to the pupil response-rate calculation). The Scottish government—within the aforementioned freedom of information response—has confirmed that this removal of 498 pupils from the Scottish sample is formed of:

- Pupils deemed ineligible (n = 356)
- Within-school exclusions (n = 142)

These figures differ slightly, however, from those provided by the OECD in the PISA 2018 technical report (2019: Table 11.2 and Table 11.8):

- Pupils deemed ineligible (n = 364)
- Within-school exclusions (n = 144)

Moreover, the OECD also provides a different figure for the sample of eligible participants (3687 versus 3767 in the Scottish national report)—as summarised in Table A1 below.

Table A1. The PISA sample for Scotland, based upon information provided by the Scottish government and the OECD

	Scottish figures	OECD figures
Sampled	4265	4195*
Ineligible	356	364
Within-school exclusions	142	144
1. Total less ineligible/withdrawn/excluded	3767	3687
Parent non-consent	122	718
Pupil absence	596	
2. Total less nonresponse	3049+	2969
3. Final stated sample	2969	2969
Unweighted response rate (2/1)	80.9%	80.5%
Unweighted response rate (3/1)	78.8%	80.5%

Notes: *This figure was not directly reported by the OECD. I have arrived at it by adding together the figure for the 'number of students sampled' provided by the OECD (3687), the number of within-school exclusions (144) and the number of ineligible pupils (364). +This figure was not directly reported by the Scottish government (2019). Rather it has been calculated by taking the stated number of eligible participants provided by the Scottish government (2019)—3767—and subtracting the numbers for parental non-consent (122) and pupil absence (596) received in response to my freedom of information request (https://www.whatdotheyknow.com/request/720228/response/1725609/attach/3/Response%20202100141438.pdf?cookie_passthrough=1).

The OECD (2019: Table 11.8) then goes on to show how, of these 3687 pupils, 718 pupils were counted as absent on the day of the test. As Table A1 illustrates, this is consistent with the figures for nonconsent/absence provided by the Scottish government in response to my freedom of information request (formed of 122 pupils whose parents did not consent and 596 pupils who were absent on the day of the test).

Unfortunately, this leads to a potentially important discrepancy in the figures reported by the Scottish government. If one takes the number of 'eligible participants' report by the Scottish government (3767) and subtracts the number of parent nonconsent (122) and absent pupils (596), one reaches a figure of 3049. Yet the final PISA 2018 sample—as reported by both the OECD and Scottish Government—is 2969.

There are therefore 80 (3049–2969) pupils that have gone missing from the Scottish sample and have not been accounted for. Yet—as Table A1 illustrates—this determines whether Scotland falls just above or just below the 80% pupil response rate threshold.

What has happened?

It appears that this discrepancy of 80 pupils is due to two individual schools having a particularly low response rate. This has led to the OECD excluding pupils within these schools from the calculation of Scotland's official pupil response rate, and including these two schools in the school nonresponse figures instead.[14]

As noted in Chapter 4 of the PISA 2018 technical report (OECD, 2019):

> A school with a student participation rate between 25% and 50% was not considered as a participating school for the purposes of calculating and documenting response rates. . .. . .. However, data from such schools were included in the database and contributed to the estimates included in the initial PISA international report.

In other words, individual schools with low levels of pupil response do not form part of the pupil response rate calculation; instead, their schools are moved to the school nonresponse figures instead. This is despite the data from such schools being included in the PISA database and contributing to a country's results. Hence the decision to include these pupils in the school nonresponse figures—rather than the pupil nonresponse calculation—is somewhat perplexing, given that 'selection' out of the study is being driven at the pupil level (i.e. their school has attempted to conduct PISA, but insufficient numbers of its pupils have agreed to take part).

How then does this issue play out in the data for Scotland?

If one downloads the international PISA database (https://www.oecd.org/pisa/data/2018database/) and looks at the data for Scotland, one sees that the number of pupils is 2998 (from across 110 schools). This is 29 more pupils (and two more schools) than the 2969 pupils (from across 108 schools) that have been used in the OECD's calculation of the official pupil response rate (see Table A1). These are presumably the 29 pupils (out of the 80 sampled) who took the PISA test in the two schools with

low pupil response rates. Indeed, it would imply that across these two schools, the pupil response rate was 36%.

If these 80 observations are included in the pupil response rate calculation—which I believe is more appropriate than treating their schools as non-respondents—then the numerator in the calculation becomes 2998 (matching the number of observations in the final PISA database), while the denominator becomes 3767. This would lead to the response rate for Scotland being 2998/3767 = 79.6%, falling below the 80% threshold.[15]

Thus, in essence, Scotland has only managed to exceed the 80% threshold—using the OECD's calculation—because two outliers (i.e. two schools with particularly low pupil response rates) have been removed from the pupil response rate calculation.

A final aside

This odd approach to calculating pupil response rates can be illustrated in two ways. First, the minimum pupil response rate a country can theoretically achieve is 50% (not 0% as one might assume). This is because any school with a pupil response rate below 50% gets moved into the school nonresponse calculations instead.

Second, it is possible for the number of pupils within sampled schools to increase, but for the 'official' pupil response rate to decrease.

To understand why, recall how the official calculation of the pupil response rate in Scotland is:

2969/3687 = 80.5%

These figures were used because the OECD decided to exclude from the calculation any school where the number of pupils tested falls between 25% and 50% of those sampled. As noted above, if these pupils are included in the pupil nonresponse figures instead, then the pupil response rate becomes:

2998/3767 = 79.6%

If we hypothetically managed to increase the response rate within these two schools up to 50% (i.e. if we managed to successfully test 40 of the 80 pupils across these two schools, rather than 29), this would increase the numerator in the 'official' pupil response rate calculation up to 2969 + 40 = 3009. Yet the denominator used in the official calculation would also increase up to 3767. This would then give an official pupil response rate of:

3009/3767 = 79.9%

In other words, we have managed to test more of our sampled pupils, yet the official pupil response rate would go *down* from 80.5% (just above the 80% threshold) to 79.9% (just below the 80% threshold). Thus Scotland could actually have got *more* of the sampled pupils to take part in PISA, but see its pupil response rate fall (with its school response rate increasing instead).

Appendix 3. Further discussion of nonresponse bias analyses.

The analysis presented in Tables 5b and 7b of the main text have been used by the NFER, OECD and national governments to justify their view that the NRBA shows samples which are 'positive' and 'representative'. Yet the comparisons upon which they focus are likely to give an overly optimistic picture of the ability of replacement schools and weighting to reduce bias in the statistic of interest (PISA scores).

To understand why, I reproduce Table 5b below for England. Column 1 provides a binary measure of historic school achievement for all of the originally sampled schools (i.e. those 199 schools that were meant to take part). Column 2 then provides the analogous figures for those 144 schools that actually did take part. This comparison reiterates the point made in the main text; lower-achieving schools were more likely to refuse to take part in the study (there were only 28% in the participating sample compared to 33% in the full original sample).

PISA has two ways of trying to compensate for this problem. The first is via the use of 'replacement schools': for those schools that refused to participate, a substitute can take its place. This is a form of imputation, and is subject to a Missing At Random (MAR) assumption. Critically, these substitute schools belong to the same stratum as the school that refused to participate. Hence a school that refused to participate and which was in the bottom 40% of the attainment distribution is replaced by another school in the bottom 40% of the attainment distribution—an entirely sensible thing to do. This, however, does mean that the figure for the 'bottom 40% of Attainment 8 distribution' can essentially only go up between Columns 2 (before replacements are included) and 3 (after replacements are included). as it has been forced to do. The extent to which this would in turn also force upwards the *unobserved* quantity of primary interest (PISA scores) is open to debate. Unless school-level PISA scores and school-level historic GCSE performance are perfectly correlated, it is likely to provide an overly optimistic picture.

The same logic then applies once the weights are applied in Column 4. The sample after replacements are included (Column 3) is still underrepresenting lower-achieving schools. The nonresponse adjustments made in the PISA weights will recognise this, and thus ensure that schools with lower historic GCSE performance are 'worth' more in the analysis. Again, this is an entirely sensible thing to do. It does, however, mean that there is a mechanical increase in the percentage of schools with low historic GCSE grades in the sample between Columns 3 and 4; the extent to which this will help to reduce the potential bias in the unobserved quantity of interest (PISA scores) is open to debate.

It thus follows that a comparison of Columns 1 and 3 and of Columns 1 and 4 in Table A2 provides an overly optimistic perspective of how well the nonresponse adjustments (replacement schools and weights) have 'worked' in reducing the likely bias in the quantity of interest (PISA scores). In my view, at best, such an analysis can only provide evidence of whether there are very serious concerns. For example, one would be extremely worried if the difference between Columns 1 and 4 continue to materially differ, even with the use of replacement schools and weighting.

Table A3 reproduces Table 7b, providing analogous figures for Northern Ireland. Focusing upon 'school type', a similar pattern emerges. Non-grammar schools were

Table A2. England

School attainment	1. All originally sampled schools	2. Participating schools before replacements included	3. Participating schools after replacements included	4. Participating schools after replacements included. Weights applied
Top 60% of Attainment 8 distribution in 2016	62%	69%	66%	62%
Bottom 40% of Attainment 8 distribution in 2016	33%	28%	31%	35%
Missing	5%	4%	3%	4%
Observations	199	144	175	175

Table A3. Northern Ireland

	1. All originally sampled schools	2. Participating schools before replacements included	3. Participating schools after replacements included	4. Participating schools after replacements included. Weights applied
School type				
Grammar	40%	46%	43%	34%
Non-grammar	60%	55%	57%	67%
Region				
Belfast	19%	24%	20%	17%
North-eastern	22%	14%	17%	17%
South-eastern	18%	14%	17%	18%
Southern	26%	29%	27%	30%
Western	17%	20%	20%	18%
School gender				
Female	13%	20%	17%	15%
Male	13%	17%	14%	15%
Mixed	75%	64%	70%	71%
N	102	66	79	79

underrepresented in the original participating sample (Column 2). Hence non-grammars were 'targeted' when including substitute schools, meaning the percentage in Column 3 increases (55% to 57%).

Once the weights are then applied in Column 4, the percentage of non-grammar schools in the sample increases further. However, what is concerning is that the adjustment (taken at face value) would seem to go too far; the percentage of non-grammar schools in Column 1 is 60% compared to 67% in the final weighted sample. This situation may have arisen because the school-level sample for Northern Ireland

is very small (just 79 schools) and this results in small cell sizes when the weights are created. However, I believe the large jump in these figures with the application of the weights is in itself a concern.

It is also important to reiterate a further point made in the main text of the paper: only a very limited number of school-level variables have been used to investigate potential bias in both England and Northern Ireland. The evidence available on potential bias, which has been presented by the NFER, English/Northern Irish governments and the OECD, is very limited (and there are doubts about the data being 'representative'). In my view, it does not provide particularly strong evidence as to whether school nonresponse has led to bias in the sample or not. Yet more detailed analyses of the data would have been possible at the time the NRBA was produced, at least in the case of England (e.g., a comparison of Key Stage 2 scores at the pupil level) but do not seem to have been conducted.

Thus, as noted in the main text of the paper, it is open to interpretation as to whether there has been much confidence in the above as evidence of the samples being 'representative' or not. What in my view is unforgivable, however, is that the NFER, OECD and the national governments have not clearly and transparently presented the evidence to allow independent individuals to make up their minds about the strength of the evidence for themselves (or even told them how this judgement was reached). Rather, they have chosen just to say that the results are 'positive' and that the data are 'representative', when—in reality—this is at best only a partial reflection of the evidence available.